doi:10.1111/j.1558-5646.2007.00080.x

ESTIMATION OF THE NUMBER OF INDIVIDUALS FOUNDING COLONIZED POPULATIONS

Eric C. Anderson[1,2] and Montgomery Slatkin[3,4]

[1]*Fisheries Ecology Division, Southwest Fisheries Science Center, 110 Shaffer Road, Santa Cruz, California 95060*

[2]*E-mail: eric.anderson@noaa.gov*

[3]*Department of Integrative Biology, University of California, Berkeley, California 94720-3140*

[4]*E-mail: slatkin@berkeley.edu*

Received June 8, 2006
Accepted December 4, 2006

A method for estimating the number of founding chromosomes in an isolated population is introduced. The method assumes that *n*/2 diploid individuals are sampled from a population and that alleles are identified at *L* unlinked loci. The population is assumed to have been founded *T* generations in the past by individuals carrying *c* chromosomes drawn randomly from a known source population, which has also been sampled. If *c* is small and the population grew rapidly after it was founded, accurate estimates of *c* can be obtained and those estimates are not sensitive to details of the history of population sizes. If *c* is larger or the population remained small after it was founded, then estimates of *c* depend on the history of population sizes. We test the performance of our method on simulated data and demonstrate its use on data from a rainbow trout (*Oncorhynchus mykiss*) population.

KEY WORDS: **Bottleneck, coalescent, importance sampling, invasions, likelihood.**

The genetic composition of a recently founded population reflects its history. Population genetic theory can be used to infer specific details of population history provided that the range of possibilities is restricted sufficiently. Here we consider the problem of estimating the number of founding chromosomes of a population that is known to have been established at a specific time in the past and that has received no immigrants afterwards. We show that, under these assumptions, accurate estimates of the number of founding chromosomes can sometimes be obtained, and we show that general properties of the neutral coalescent model in a population of variable size can indicate whether accurate estimates can be obtained in principle.

Estimating the number of founding chromosomes of an isolated population can allow tests of specific hypotheses about the history of a population. One could ask, for example, whether the current genetic composition of a population is consistent with historical information. Estimating the number of founding chromosomes may also be useful for understanding the intensity of founder effects, which are widely invoked but rarely tested for. Wright (1931)'s shifting balance theory, Mayr (1954)'s theory of genetic revolutions, and various theories of speciation (Carson and Templeton 1984) all assume that substantial genetic changes occur when populations are founded by small numbers of individuals. In human genetics, founder effects are often assumed to account for the presence of Mendelian diseases found in unusually high frequencies in isolated populations (Vogel and Motulsky 1996), but, at present, tests of founder effects focus on disease-associated alleles rather than on patterns of genetic variation at other loci (Risch et al. 2003; Slatkin 2004).

The problem we address here is closely related to the problem of detecting whether an isolated population has experienced an extreme reduction (a "bottleneck") in population size. Nei et al. (1975) were the first to explore quantitatively the effects of bottlenecks on genetic diversity. They showed that the reduction in heterozygosity but not the reduction in the number of alleles is predicted by Wright (1938)'s effective population

size. They also noted that a bottleneck resulted in a skewed frequency spectrum, with a lower proportion of low frequency alleles than in a population of constant size. Nei et al. (1975) argued that the reduced variability of allozyme loci found in the Bogata population of *Drosophila pseudoobscura* resulted from a founder event.

More recently, Luikart et al. (1998a,b), and Beaumont (1999) have developed statistical tests of whether bottlenecks have occurred. Those tests are based on detecting differences between an observed allele frequency spectrum and the spectrum expected in a population of constant size. Luikart et al. (1999) also propose a method-of-moments estimator that uses two temporally spaced genetic samples to detect bottlenecks that occur in the interval between the samples. Our analysis differs in two ways from that of Luikart et al. (1998a,b) and Beaumont (1999). First, we assume that a founder event occurs at a known time in the past, whereas Luikart et al. (1998a) and Beaumont (1999) test whether a bottleneck occurred at any time in the past. Second, we assume that samples are available from the source population. Our method differs from that of Luikart et al. (1999) because ours is a maximum likelihood method based on the coalescent, and is not a method that relies on the change over time solely of the variance in allele frequencies. Our method can be used to test for the occurrence of a bottleneck at the time the population was founded by testing the hypothesis that the number of founding chromosomes did not differ significantly from twice the current population size.

In the following, we first identify conditions under which it is feasible to estimate the number of founding chromosomes. Then we describe the model and calculations that allow maximum-likelihood estimation. Finally we test the method against simulated data and then illustrate its use by applying it to data from a rainbow trout *Oncorhynchus mykiss* population.

FEASIBILITY OF ESTIMATING THE NUMBER OF FOUNDING CHROMOSOMES

In this section, we use general properties of the neutral coalescent (Tavaré 1984) to determine whether it is possible in principle to estimate the number of founding chromosomes of an isolated population. In some situations, it will be impossible to estimate the number of founding chromosomes with confidence, even if sufficient genetic data were available to allow accurate estimation of the number of ancestral lineages present at the time the population was founded, because that number would be expected to be much less than the number of founding chromosomes and the difference would depend on details of the history of population sizes that are probably unknown.

We assume that a population, which we refer to as the "colony," was established by $c/2$ diploid migrants from the "source" population T generations in the past, and that the population size in the colony between T generations in the past and the present ($t = 0$) are known, $N(t)$. (Table 1 provides a guide to the mathematical notation used in this section and the next.)

Table 1. Mathematical notation used to describe the model.

$A(t)$	Number of lineages at time t ancestral to the n sampled colony chromosomes
$A_S(t)$	Number of lineages at time t ancestral to the n_S sampled source chromosomes
c	Number of chromosomes present amongst the colony founders at time T
K	Number of alleles observed in the samples from the colony and the source
n	Number of chromosomes sampled from the colony population at $t = 0$
n_S	Number of chromosomes sampled from the source population at $t = 0$
$N(t)$	Number of diploids in the colony population at time t
N_K	Carrying capacity of the colony population
r	Intrinsic rate of growth of the colony population
t	A variable that indicates time in generations. Varies from 0 (the present) to T
T	The number of generations in the past that the colonization occurred
τ_S	T generations scaled by the size of the source population
$\boldsymbol{x}_0$	vector of allelic counts observed in the n genes from the colony
$\boldsymbol{x}_T$	unobserved allelic counts among the colony's $A(T)$ ancestral lineages
$\boldsymbol{y}_0$	vector of allelic counts observed in the n_S genes from the source
$\boldsymbol{y}_T$	unobserved allelic counts among the source's $A_S(T)$ ancestral lineages

We will assume that a sample of $n/2$ individuals is taken at the present time. We will be concerned with the ancestry of a single locus and assume that the probability distribution obtained for a single locus represents the distribution across the unlinked loci surveyed. The number of ancestral lineages at any time in the past can be found under neutrality by using coalescent theory. Let $A(t)$ be the random variable representing the number of ancestral lineages in generation t in the past. Given $A(t)$ and the population size in the preceding generation, $N(t+1)$, the distribution of $A(t+1)$ is the probability that, if $A(t)$ balls are randomly distributed into $2N(t+1)$ boxes, $A(t+1)$ boxes are nonempty (Kingman 1982)

$$\Pr(A(t+1) = k | N(t+1), A(t) = a) = \binom{2N(t+1)}{k} \sum_{v=0}^{k} (-1)^{k-v} \binom{k}{v} \left(\frac{v}{2N(t+1)} \right)^{a}. \quad (1)$$

This model provides the transition probabilities of a Markov chain on the state space $A(t) = 1, \ldots, n$. The initial condition is

$A(0) = n$ and there is a single absorbing state at $A(t) = 1$. It is a pure death process, meaning that $A(t+1) \leq A(t)$. In most applications of coalescent theory, it is assumed that $N(t)$ is sufficiently large and n is sufficiently small that the probability that $A(t)$ decreases by more than 1 in a single generation is vanishingly small, which we will refer to as the "diffusion limit." In the diffusion limit,

$$\Pr(A(t+1) = A(t)) = 1 - \binom{A(t)}{2}/(2N(t+1)) \qquad (2)$$

$$\Pr(A(t+1) = A(t)-1) = \binom{A(t)}{2}/(2N(t+1) \qquad (3)$$

and the approximate distribution of $A(T)$ for a given history of population sizes, $N(t)$, $0 \leq t \leq T$, and given sample size, n can be found in closed form (Tavaré 1984):

$$P(A(T) = j|n, N(t)) = \begin{cases} \sum_{k=j}^{n} \frac{(-1)^{k-j}(2k-1)j_{(k-1)}n_{[k]}}{j!(k-j)!n_{(k)}} \\ \quad \times \exp\{-k(k-1)\tau/2\}, 2 \leq j \leq n \\ 1 - \sum_{k=2}^{i} \frac{(-1)^{k-j}(2k-1)j_{(k-1)}}{j!(k-j)!} \\ \quad \times \exp\{-k(k-1)\tau/2\}, j = 1 \end{cases} \qquad (4)$$

where $i_{[k]} = i(i-1)(i-k+1)$ and $i_{(k)} = i(i+1)(i+k-1)$, and τ is the number of generations scaled by population size, $\tau = \sum_{t=1}^{T} \frac{1}{2N(t)}$ (Griffiths and Tavaré 1994).

Because we will allow the possibility of very small initial population sizes, possibly as small as $c = 4$ representing the founding of a population by a single female singly inseminated, we will not assume the diffusion limit applies. In that case, it appears that the distribution of $A(t)$ cannot be expressed in closed form and instead must be obtained either by simulation or by an exact iteration of the Markov chain. For a given history of population sizes, and given sample size, the distribution of $A(T)$ will be written as $P(A(T)|n, N(t))$, with the dependence on n and $N(t)$ omitted unless needed for clarity.

Our concern is with estimating c from the genetic composition of the L loci surveyed. The alleles found in the sample at the present time are the alleles present on the founding chromosomes plus any that arose by mutation since founding. Therefore, the genetic composition of the sample is determined not by c but by $A(T)$, because only that number of founding lineages is represented in the sample. There are two possibilities. If there is a high probability that $A(T)$ is close to c under a wide range of feasible demographic histories, then it is reasonable to assume that the method described in the next section will lead to an estimate of c, because all or nearly all founding chromosomes are represented in the sample. If, on the other hand, $A(T) \ll c$ with high probability, then our ability to estimate $c = 2N(T)$ depends on the relationship between $A(T)$ and the specific model of demographic history through the dependence of $P(A(T)|n, N(t))$ on $N(t)$. Because the true demographic history of an isolated population is probably not well known, only in the first case can we reasonably expect to obtain an accurate estimate of c even if very extensive genetic data were available. In the second case, the best that can be done in practice is that $A(T)$ can be estimated and the relationship between c and $A(T)$ examined.

To illustrate the dependence of $A(T)$ on c and the demographic history, we assumed that the population size followed a logistic curve with intrinsic rate of increase r, carrying capacity N_K, and initial size, $c/2$

$$N(t) = \frac{ce^{r(T-t)}}{2 + c(e^{r(T-t)} - 1)/N_K}. \qquad (5)$$

Under this model, it is straightforward to simulate the ancestral process for given r, c, N_K, T, and n to obtain an approximation for $P(A(T))$. We also consider an extreme case of very large r because that results in the most extreme distribution of $A(T)$ and one that can be calculated analytically. This extreme distribution, which we will denote by $P_x(A(T))$, is obtained by computing the distribution of $A(T-1)$ from (4) and then using (1) to model the random assignment of $A(T-1)$ lineages to c chromosomes.

To illustrate our results, we estimated $P(A(T))$ for various parameter values of our model. In all cases, $N_K = 1000$ and each curve shown summarizes the results of 1000 replicates. Figure 1 shows the history of population sizes under our model for the case with $c = 10$ and $T = 50$. Figure 2 shows $P(A(T))$ for $T = 20$ and

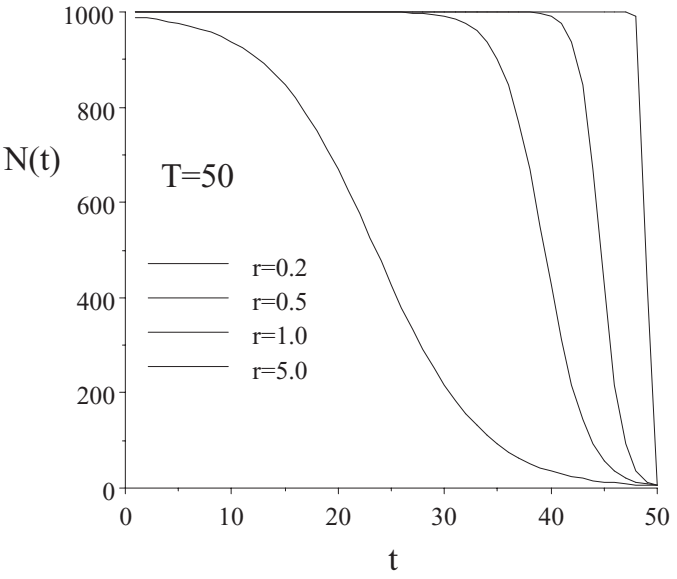

Figure 1. Population size as a function of time, computed using the logistic growth model of equation 5. Note that points that are further right on the *x*-axis are further back in time.

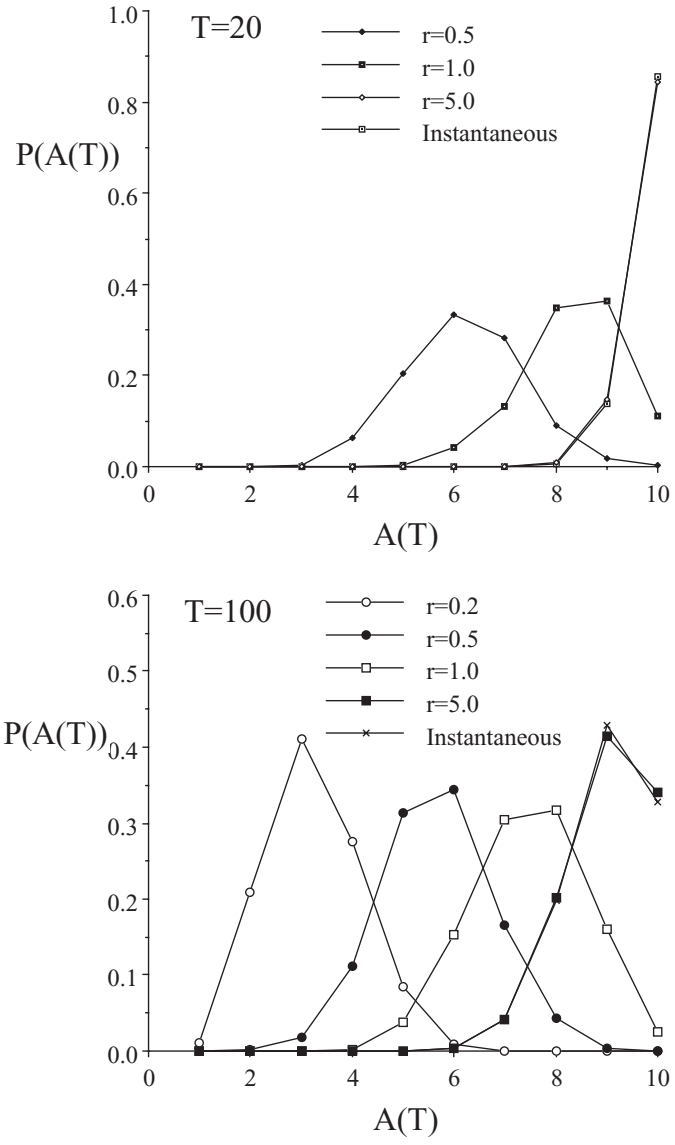

Figure 2. Probability distribution of $A(T)$, the number of lineages remaining after T generations given a population size starting with $N(T) = 5$ (i.e., $c = 10$) and growing via equation 5 to a carrying capacity of $N_K = 1000$. $P(A(T))$ was approximated by simulation.

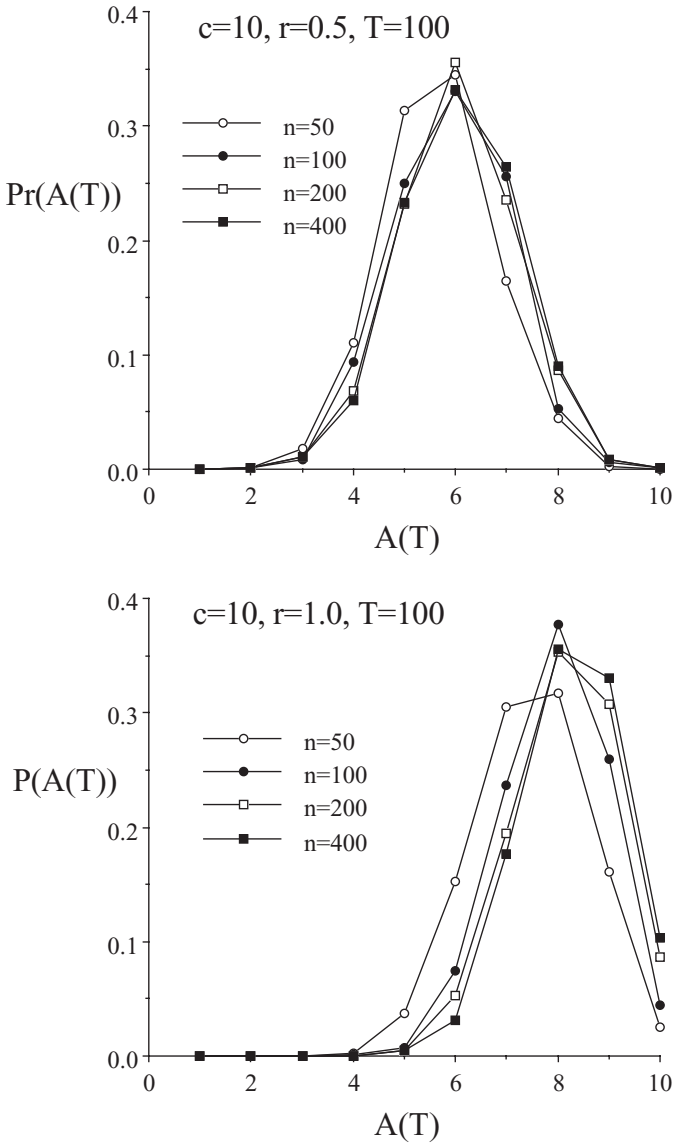

Figure 3. The influence of sample size n on $A(T)$.

$T = 100$ with $c = 10$. In both cases, the instantaneous approximation derived above provides an excellent approximation when $r = 5$, for which $N(t)$ increases from $c/2$ to N_K in three generations. For smaller and more biologically reasonable r, the instantaneous approximation is not adequate, implying that the slower population growth results in a substantial number of additional coalescent events that reduce the number of ancestral lineages. For the smaller values of r, $A(T)$ is typically less then c, implying that even perfect knowledge of the distribution of $A(T)$ would not lead to an estimate of c unless the logistic model were accurate.

Figure 3 shows that increasing the sample size (n) can increase $A(T)$ slightly but not necessarily by much. The reason is that the initial rate of coalescence is proportional to n^2. Increasing n results only in an increase in the number of coalescence events occurring in the most recent few generations, rather than an increase in $A(T)$. Figure 4 illustrates that, for given parameter values, larger values of c result in proportionately fewer ancestral chromosomes being represented in the sample.

In summary, $A(T)$ is expected to be less than c for all but very high—and possibly biologically unreasonable—intrinsic rates of growth. However, as shown in Figure 4, with a biologically reasonable intrinsic growth rate such as $r = 1.0$, the number of founding chromosomes, c, has a considerable effect on the number of ancestral lineages $A(T)$. This relationship can be exploited to estimate c using genetic data, given an assumed or known population history.

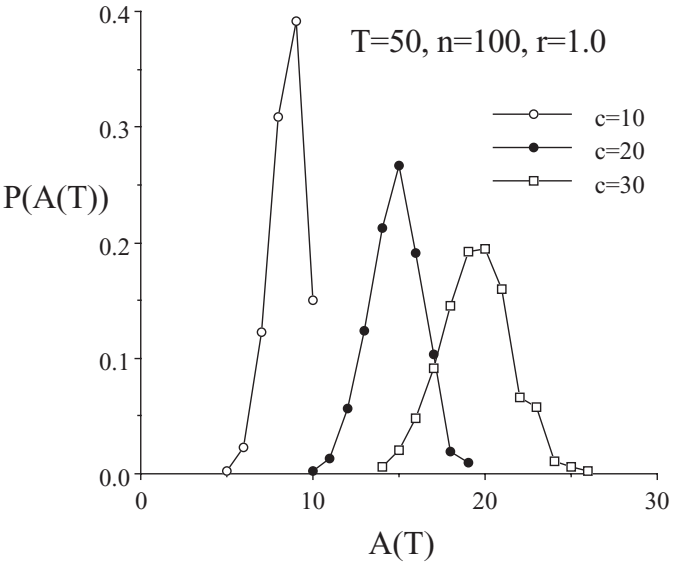

Figure 4. The influence of *c* on the distribution of *A(T)*.

Estimation of c from Genetic Data

As before, we refer to the recently founded population as the "colonized" population or the "colony" and we refer to the population from which the colonizers originated as the "source" population. In this section we describe a method to compute the likelihood for $N(t)$—the colony's population size history—given samples of polymorphic genetic markers taken in the present from the colonized and source populations. Because the number of founding chromosomes is $c = 2N(T)$, this likelihood can be used to estimate c. We assume that there is no mutation. Such an assumption is reasonable when few generations have elapsed since the time of colony founding and/or the mutation rates of the genetic markers are not high.

We establish the notation and the likelihood model in the context of a single locus at which K alleles are observed in the combined genetic samples from the colony and source. The likelihood for multiple, independently segregating loci that are not in linkage disequilibrium in the source population at time T is simply the product over loci of the single-locus likelihoods. As before, n gene copies are sampled from the colonized population, and we let n_S denote the number sampled in the present day from the source. The vectors $\boldsymbol{x}_0 = (x_{0,1}, \ldots, x_{0,K})$ and $\boldsymbol{y}_0 = (y_{0,1}, \ldots, y_{0,K})$ denote the numbers of the K different alleles in the present-day samples from the colonized and source populations, respectively; $n = \sum_{k=1}^{K} x_{0,k}$ and $n_S = \sum_{k=1}^{K} y_{0,k}$.

The n genes from the colony descended from $A(T)$ ancestral lineages extant at time T, and the n_S genes from the source descended from $A_S(T)$ ancestral lineages. Both $A(T)$ and $A_S(T)$ are unknown, as are the allelic types of those ancestral lineages, denoted $\boldsymbol{x}_T = (x_{T,1}, \ldots, x_{T,K})$ and $\boldsymbol{y}_T = (y_{T,1}, \ldots, y_{T,K})$, respectively. However, these variables are included as latent variables in the likelihood model. Finally, the allele frequencies in the source population at the time of colonization are additional latent variables in the model which we denote by $\boldsymbol{p} = (p_1, \ldots, p_K, p_{K+1})$ where p_{K+1} is the frequency in the source population at time T of all alleles that were not detected in the samples from the colony or the source. Omitting the alleles in the $K + 1$ category simply changes the likelihood by a constant factor, which does not alter inferences made using the likelihood, so we redefine $\boldsymbol{p}$ to be the vector $(p_1, \ldots, p_K)$ with $\sum_{k=1}^{K} p_k = 1$.

Recall that $P(A(T)|n, N(t))$ denotes the marginal probability (unconditional on any genetic data) that n gene copies sampled from the colony at time 0 descended from $A(T)$ ancestral lineages at the time of founding, conditional on the population size history $N(t)$. We will assume that the source population is large enough so that the diffusion limit applies and the distribution of the number of ancestral lineages in the source population, $P(A_S(T)|n_S, \tau_S)$, is given by (4) with an appropriately scaled time τ_S. These probabilities may be combined with probabilities of the observed and latent variables described above to derive the likelihood for $N(t)$. To achieve this, we first derive the joint probability of all the variables, and then integrate out the latent variables. The joint probability of the latent and observed variables is:

$$\begin{aligned} &P(\boldsymbol{x}_0, \boldsymbol{x}_T, \boldsymbol{y}_0, \boldsymbol{y}_T, A(T), A_S(T)|\boldsymbol{p}, n, n_S, N(t), \tau_S) \\ &\quad = P(\boldsymbol{x}_0|\boldsymbol{x}_T, A(T), n)P(\boldsymbol{y}_0|\boldsymbol{y}_T, A_S(T), n_S) \\ &\qquad \times P(\boldsymbol{x}_T|A(T), \boldsymbol{p})P(\boldsymbol{y}_T|A_S(T), \boldsymbol{p}) \\ &\qquad \times P(A(T)|n, N(t))P(A_S(T)|n_S, \tau_S) \end{aligned} \tag{6}$$

In words, the joint probability is the product of six conditional probabilities: (1) the probability that n genes of allelic type $\boldsymbol{x}_0$ descended from $A(T)$ ancestral lineages having allelic types according to $\boldsymbol{x}_T$; (2) the probability that n_S genes of allelic type $\boldsymbol{y}_0$ descended from $A_S(T)$ ancestral lineages having allelic types according to $\boldsymbol{y}_T$; (3) the probability that $A(T)$ genes of allelic types according to $\boldsymbol{x}_T$ are drawn from a large population in which the allele frequencies are $\boldsymbol{p}$; (4) the probability that $A_S(T)$ genes of allelic types according to $\boldsymbol{y}_T$ are drawn from a large population in which the allele frequencies are $\boldsymbol{p}$; (5) the probability that n lineages coalesce into $A(T)$ lineages given the population history $N(t)$; and finally (6) the probability that n_S lineages coalesce into $A_S(T)$ lineages in scaled time τ_S.

In the diffusion limit and in the absence of mutation, the allelic types carried by genes descended from ancestral lineages possessing certain allelic types follows a form of the Dirichlet-compound multinomial distribution (Hoppe 1984). Thus the distribution of allelic types in the sample from the source population can be written as

$$P(\boldsymbol{y}_0|\boldsymbol{y}_T, A_S(T), n_S) = \binom{n_S - 1}{A_S(T) - 1}^{-1} \prod_{k=1}^{K} \binom{y_{0,k} - 1}{y_{T,k} - 1}, \tag{7}$$

where $y_{0,k} \geq y_{T,k}\ \forall k$, and where we define the binomial coefficient $\binom{-1}{-1}$ to be 1 (for the case that $y_{0,k} = y_{T,k} = 0$). Such a distribution technically holds only in the diffusion limit in which no more than two genes coalesce in any coalescent event. This might not be the case in a small colonized population shortly after founding; however, we also use this distribution to describe $P(\boldsymbol{x}_0 | \boldsymbol{x}_T, A(T), n)$, recognizing that it is an approximation. As will be seen in Simulated Data, this approximation does not seem to bias the estimation of c, even when c is very small. Additionally, we experimented with a more elaborate model to account for the increased variance in the number of descendants per lineage that occurs when the diffusion limit does not hold. We found the more elaborate model failed to outperform the simpler model, so we used the simpler model, (7).

$P(\boldsymbol{x}_T | A(T), \boldsymbol{p})$ and $P(\boldsymbol{y}_T | A_S(T), \boldsymbol{p})$ both follow the multinomial distribution. In each case, we assume that the allelic types of the ancestral lineages are drawn with replacement from the allele frequencies in the source population. For $\boldsymbol{x}_T$:

$$P(\boldsymbol{x}_T | A(T), \boldsymbol{p}) = A(T)! \prod_{k=1}^{K} \frac{p_k^{x_{T,k}}}{x_{T,k}!}. \quad (8)$$

The distribution for $\boldsymbol{y}_T$ is identical with y's replacing the x's and $A_S(T)$ replacing $A(T)$.

The likelihood function for $N(t)$ (and hence c, Because $N(T) \equiv c/2$) given only the observed variables is proportional to

$$P(\boldsymbol{x}_0, \boldsymbol{y}_0 | \boldsymbol{p}, n, n_S, N(t), \tau_S)$$

considered as a function of $N(t)$. To obtain this we must sum (6) over $A(T)$, $A_S(T)$, $\boldsymbol{x}_T$, and $\boldsymbol{y}_T$. We also must integrate over all values of the nuisance parameter $\boldsymbol{p}$ (hence considering an integrated likelihood). This requires that we assign a prior distribution, $P(\boldsymbol{p})$ to $\boldsymbol{p}$. For this prior we use the Dirichlet density with parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$. Typically, each λ_i will be $1/K$ or 1, providing the "unit-information" or the uniform prior, respectively. The Dirichlet distribution is the equilibrium distribution of a K-allele model with reversible mutation, and is also the asymptotic allele frequency distribution for a population in drift-migration equilibrium (Wright 1937). Being the conjugate prior of the multinomial distribution, it also has desirable mathematical properties that we exploit later.

The likelihood is thus

$$L(N(t)) \propto \sum_{A(T)=K'}^{n} \sum_{A_S(T)=K_C}^{n_S} \sum_{\boldsymbol{x}_T} \sum_{\boldsymbol{y}_T} \int_{\boldsymbol{p}} P(\boldsymbol{x}_0, \boldsymbol{x}_T, \boldsymbol{y}_0, \boldsymbol{y}_T, A(T), A_S(T) | \boldsymbol{p}, n, n_S, N(t), \tau_S) P(\boldsymbol{p}) d\boldsymbol{p} \quad (9)$$

where K' is the number of alleles appearing only in the sample from the colony and K_C is the number of alleles appearing only in the sample from the source population. The sums over $\boldsymbol{x}_T$ and $\boldsymbol{y}_T$ can have many terms in them, especially if the sample sizes are large and the number of alleles is more than five or six. This makes it intractable to evaluate the sums and integral in (10) directly—some approximation is necessary. Sums similar to this have appeared in other contexts; for example, in the likelihood of admixture proportions in recently admixed populations. Chikhi et al. (2001) developed an MCMC method for approximating the sums, but report that it required almost a week of computer time to run their algorithm.

We investigate a simple approximation—that of assuming no genetic drift occurred in the source population between time T and 0. Making such an assumption reduces the computational burden so that the only difficult task that remains is the sum over values of $\boldsymbol{x}_T$. This is not, in itself, an easy problem; however, a fast importance sampling algorithm for approximating the sum was introduced in Anderson (2005) for the purpose of estimating N_e from two temporally spaced samples. If allele frequencies are assumed to be the same in the source population at time T and time 0, then the sample from the source at time 0 can be treated as a sample from time T, and the probability model becomes similar to that in the N_e estimation problem. Specifically, the calculation described in equation 13 in Anderson (2005) is identical to that of computing $L(A(T)|\boldsymbol{x}_0, \boldsymbol{y}_0)$, the likelihood of $A(T)$ given the genetic data. To compute $L(N(t))$ for a single locus, one first uses the importance sampling algorithm to compute $L(A(T) = a|\boldsymbol{x}_0, \boldsymbol{y}_0)$ for $K' \leq a \leq n$. Then for any history of colonized population sizes, $N(t)$, we have

$$L(N(t)) = \sum_{a=K'}^{n} L(A(T) = a|\boldsymbol{x}_0, \boldsymbol{y}_0) P(A(T)|n, N(t)). \quad (10)$$

This calculation, as well as the importance sampling algorithm, are implemented in the computer package *nfcone* that we used to test the method on simulated data as described below.

Simulated Data

We simulated data under two scenarios that illustrate the general behavior of this estimation method. In the "Large Population" scenario, the source population was of constant size with $N_e = 5000$ diploid organisms, and the carrying capacity of the colonized population was $N_K = 3000$ diploids. The colony was founded by c chromosomes, 500 generations before the present, and the intrinsic rate of growth of the colony was r. Genetic data were assumed to consist of 12 independently segregating loci taken from 100 diploids of the colony and 100 of the source population. Each locus was assumed to have 12 alleles in the source population at the time of founding. The allele frequencies at each locus were randomly simulated from a Dirichlet (2, 2, ..., 2) distribution. The simulations of the "Small Population" scenario were identical except that the source population was of size $N_e = 1000$, the

carrying capacity of the colony was set at $N_K = 300$, and the time of colonization was set at $T = 25$ generations.

Five hundred independent replicate data sets were simulated for all combinations of $c \in \{4, 8, 12, 18, 28, 40, 60, 80\}$ and $r \in \{0.5, 1.5, 4.0\}$. For the data analysis, it was assumed that the carrying capacity of the colony, the effective size of the source population, and the time of colony founding were known without error; however, the data were analyzed under a number of different assumptions about the intrinsic rate of increase. For each replicate data set, $L(A(T)|\boldsymbol{x}_0, \boldsymbol{y}_0)$ was computed using the importance sampling algorithm in *nfcone* and then $L(N(t))$ for values of $c \in \{2, 4, \ldots, 120\}$ was computed for each value of $r \in \{0.1, 0.25, 0.5, 0.75, 1.0, 1.5, 2.0, 4.0\}$. For each combination of c and r, $P(A(T)|N(t), n)$ was approximated by simulation using 20000 replicates. It is important to realize that we did not try to *jointly* estimate the number of founding chromosomes and the intrinsic rate of increase of each population—it is likely not possible to do so accurately.

Our estimator for the number of founding chromosomes behaves remarkably well in a statistical sense. If the analysis is performed assuming the correct intrinsic rate of increase, the estimator appears to be unbiased. For each true value of c, the mean value of the 500 maximum-likelihood estimates (MLEs) was very close to the true value in both the Large and Small Population simulations (Figs. 5(A–C) and 6(A–C)). For larger values of c ($\geq$60), and especially in the Large Population simulations with high values of r (1.5 and 4.0), it appears that the estimator for c may be downward biased. However, this likely results from the fact that the highest value of c we considered in the estimation procedure was 120. Had we allowed values of c larger than 120 in our estimation procedure, the estimator would likely also be unbiased, or nearly so, for values of $c \geq 60$.

Panels A–C in Figures 5 and 6 also show the extent of the standard deviation of the MLEs for c. It is clear that if the true value of c is small (< 30), and the intrinsic rate of increase is known precisely, then our method allows precise estimation of c.

Unfortunately, as noted previously, the estimate of c may be quite sensitive to the value of the intrinsic rate of growth assumed. This is confirmed in panels D–F of Figures 5 and 6. It is apparent that assuming an r that is less than the true value leads to overestimates of c, wheras assuming an r that is greater than the true value leads to underestimates of c. More encouraging, however, it is also evident that for $r = 1.5$—a biologically reasonable intrinsic rate of growth for some species—the error associated with assuming a value of r greater than the true value is not extreme, especially if the true value of c is low. This result reflects the finding in Feasibility of Estimating the Number of Founding Chromosomes that, if c is small enough and r is high enough, then c will be close to $A(T)$ and c can be estimated without accurately knowing the true intrinsic rate of increase of the population.

In our estimation procedure, we have made the assumption that there has been no genetic drift in the source population since

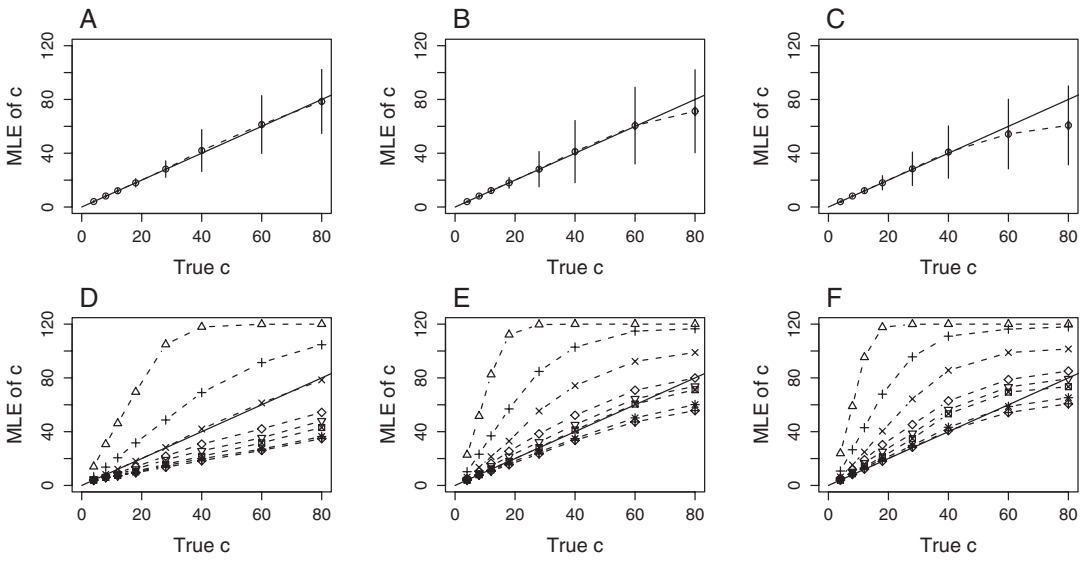

Figure 5. **Summary of simulation results for the Large Population scenario. Simulation conditions are described in the text. The top three panels (A–C) show the results when data are analyzed assuming the true growth rate *r*. Open circles represent the mean MLE from 500 simulations and the vertical bars represent the standard deviation of the MLEs. True value of *r* increases from left to right: in *A*, $r = 0.5$; in *B*, $r = 1.5$; and in *C*, $r = 4.0$. The bottom three panels show the mean MLE from 500 replicates under the assumption of a range of *r* values. True *r* increases from left to right: in *D*, $r = 0.5$; in *E*, $r = 1.5$; and in *F*, $r = 4.0$. The value of *r* assumed for the analysis is denoted by the different symbols in the plots: $\triangle = 0.1$; $+ = 0.25$; $\times = 0.5$; $\diamond = 0.75$; $\triangledown = 1.0$; $\boxtimes = 1.5$; $* = 2.0$; $\oplus = 4.0$.**

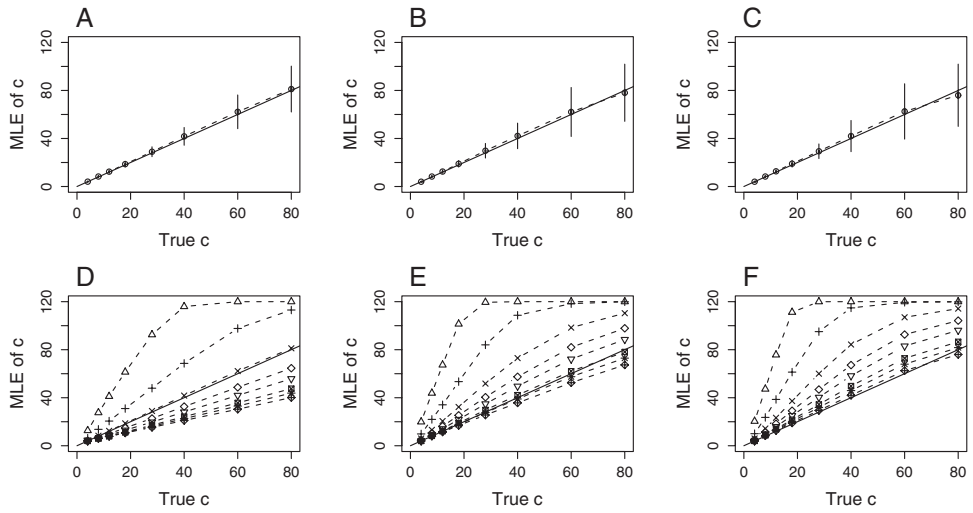

Figure 6. Summary of simulation results for the Small Population scenario. Simulation conditions are described in the text. See caption of Figure 5 for the explanation.

the time of colony founding. However, in the simulations, we performed, the source population was of finite size and some genetic drift did occur. In the Large Population simulations the amount of genetic drift can be characterized by $\tau_S = T/2N_e = 0.05$ and for the Small Population simulations $\tau_S = 0.0125$. Although the assumption of no genetic drift in the source population does not seem to bias the MLE of c, it may lead us to underestimate the uncertainty in the model, and hence overestimate the precision of the MLE. We investigated this by constructing approximate 95% confidence intervals for the estimates of c using the two-units support limit (Edwards 1992); that is, the low endpoint of the interval was the lowest value of c for which the log likelihood was within two of the maximum likelihood, and the high endpoint was the highest value of c having a log likelihood within two of the maximum. In Table 2 we list the percentage of replicates in which the interval did not contain the true value of c. In the Small Population simulations, the true value was contained in the confidence interval close to 95% of the time over all the simulation conditions. However, in the Large Population simulations, in which more drift is expected to have occurred in the source population, the true value of c was contained in the confidence intervals less than 95% of the time, indicating that when more genetic drift is expected to occur in the source population, the approximation of no drift may negatively impact the inference.

Trout Dataset

The Scott Creek drainage (Santa Cruz County, California) is inhabited by rainbow trout (*O. mykiss*) that exist both in an anadromous form that matures in the ocean, but returns to fresh water to spawn, and a resident form, whose entire life history takes place in fresh water. Big Creek, a tributary, travels over a roughly 30 m waterfall, impassable to anadromous *O. mykiss*, several kilometers above its confluence with Scott Creek. Above this waterfall is a population of resident *O. mykiss* of uncertain origin. Some contend that the above-falls reach was colonized long ago by anadromous *O. mykiss* before the geomorphic changes occurred, which now prevent access to the above-falls reach. A different hypothesis suggests that fish in the above-falls reach are the descendants of juveniles derived from the downstream anadromous population that were transported by early foresters above the falls in buckets. The landowners' family journals refer to such transplants occurring in 1906 (S. Hayes, pers. comm.).

Between 2002 and 2005, nonlethal fin-clips were obtained from 297 adult, anadromous *O. mykiss* below the falls and from 166 *O. mykiss* of mixed ages above the falls. DNA was extracted from these fin clips and amplified using the polymerase chain reaction to yield the genotypes at 18 microsatellite loci for each fish. The number of alleles observed among both populations varied from three to 33 between loci. Here, we use the methods developed in the previous sections to estimate the number of individuals transported above the falls, assuming the hypothesis that the above-falls population was derived exclusively from transplants of young, anadromous *O. mykiss* in 1906. We are thus designating the anadromous population as the source population and the above-falls population as the "colony," and assuming genetic drift in the anadromous population has been negligible compared to that in the colony. The estimated number of individuals transported above the falls will be one half the estimated number of founding chromosomes, because these markers have diploid inheritance in *O. mykiss*.

We begin by computing, for each locus, $L(A(T) = a|\boldsymbol{x}_0, \boldsymbol{y}_0)$, the likelihood of the number of lineages remaining at the time of founding, given the samples collected from both the source

Table 2. Percentage of two-unit support-limit confidence intervals containing the true value of c. All estimates were made assuming the true value r of the intrinsic rate of increase. %Below, %In, and %Above are the percentages of 500 replicate simulations in which the true value was below, within, or above the confidence interval, respectively. Values for the Large Population simulation, which were biased due to evaluating the likelihood only to $c = 120$, are omitted.

c	r	Small population			Large population		
		% Below	% In	% Above	% Below	% In	% Above
4	.5	.4	99.6	.0	.4	99.6	.0
4	1.5	.0	100.0	.0	.0	100.0	.0
4	4.0	.0	100.0	.0	.0	100.0	.0
8	.5	2.2	95.6	2.2	2.0	93.8	4.2
8	1.5	4.4	94.0	1.6	3.2	93.6	3.2
8	4.0	5.6	93.2	1.2	3.2	94.2	2.6
12	.5	3.2	94.6	2.2	2.4	95.6	2.0
12	1.5	3.2	94.8	2.0	4.2	91.4	4.4
12	4.0	6.2	92.4	1.4	3.8	91.2	5.0
18	.5	3.0	95.4	1.6	2.4	93.6	4.0
18	1.5	6.0	91.6	2.4	3.2	90.2	6.6
18	4.0	6.4	92.2	1.4	2.2	92.0	5.8
28	.5	4.2	93.6	2.2	2.6	90.4	7.0
28	1.5	6.2	91.0	2.8	1.8	87.0	11.2
28	4.0	4.0	93.0	3.0	1.6	88.2	10.2
40	.5	3.8	93.0	3.2	2.2	91.2	6.6
40	1.5	3.8	94.4	1.8	1.0	87.4	11.6
40	4.0	3.0	93.6	3.4	1.2	90.6	8.2
60	.5	3.2	94.2	2.6	.4	92.2	7.4
60	1.5	2.6	91.4	6.0	.4	95.6	4.0
60	4.0	1.6	93.0	5.4	–	–	–
80	.5	2.4	94.8	2.8	.4	99.0	.6
80	1.5	1.0	94.8	4.2	–	–	–
80	4.0	.0	93.4	6.6	–	–	–

and the colony. This quantity, which depends only on the genetic data, and not on the assumed length of time since the founding event, can be calculated using the software *nfcone* (full details of the implementation of these calculations are distributed with the software). The curves of log $L(A(T) = a|\boldsymbol{x}_0, \boldsymbol{y}_0)$ are shown in Figure 7. It is clear that the maximum of $L(A(T) = a|\boldsymbol{x}_0, \boldsymbol{y}_0)$ occurs most often with a between 25 and 75, with some loci showing peaks falling outside that range. To use those values of $L(A(T) = a|\boldsymbol{x}_0, \boldsymbol{y}_0)$ in (10) to estimate the number of founding chromosomes, it is necessary to compute $P(A(T)|n, N(t))$—the probability of having a lineages remaining given that the colony has had population sizes of $N(t)$ between the time of founding and the time of sampling.

In this case, there is no record of population sizes above the falls. From electro-fishing surveys, however, the population

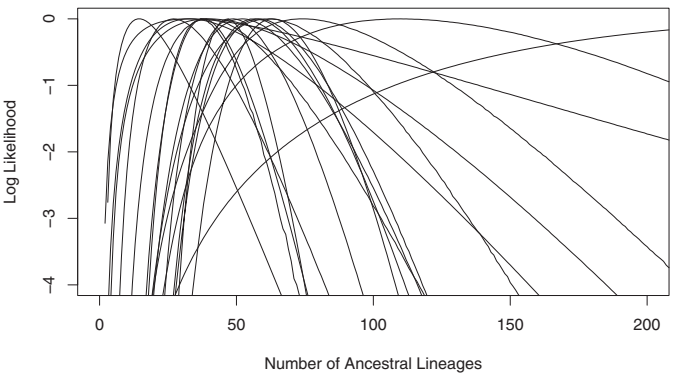

Figure 7. $L(A(T) = a|\boldsymbol{x}_0, \boldsymbol{y}_0)$ plotted as a function of a, the number of remaining lineages, ancestral to the sample from the above-falls *O. mykiss* population, at the time of colony founding. Each curve represents the log likelihood for a single locus, shifted as necessary so that its maximum value is 0.

size is estimated today to be about 1000 trout (S. Hayes, pers. comm.). We use that figure as a carrying capacity and, using the software program *spip* (Anderson and Dunham 2005), simulate an age-structured population of individuals that grows from $c/2$ individuals (67% of which are one-year-olds, 20% two-year-olds, and 13% three-year-olds) to 1000 individuals. Individuals are sampled from this simulated population, and the ancestry of their genes is simulated upward through their pedigree back to the gene copies carried by the founders of the colony. This constitutes a single replicate simulation of the number of lineages ancestral to the sample from the colony. This procedure was repeated 3000 times for each value of $c/2$ in the set $\{20, 40, 60, 80, 120, 160, 200, 260, 320, 380, 480, 600\}$, giving a Monte Carlo approximation to the distribution $P(A(T)|n, N(t))$. Note that each value of the vector of population sizes through time, $N(t)$, is indexed by a value of $c/2$.

Reproduction and survival in the simulated age-structured population were governed by the following Leslie matrix:

$$\mathbf{A} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\ \begin{pmatrix} 0 & 0 & w & 2w & 3w & 4w & 5w & 6w \\ .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & .7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & .8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .9 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & .9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & .7 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .6 & 0 \end{pmatrix} \end{array}. \quad (11)$$

In text, 50% of one-year-olds survive to be two-year-olds, 70% of two-year-olds survive to be three-year-olds, and so forth. No individual lives past eight years. Females of ages one and two do not reproduce, however, each three-year-old female produces, on average, w female offspring per year ($2w$ offspring total,

assuming an equal sex ratio) that survive to age one. Each four-year-old is expected to produce $2w$ female offspring per year surviving to age one, and so forth. The increasing number of offspring reflects the greater fecundity of older, larger females. Variance in reproductive success of males and females was set so that in the absence of any population size fluctuations the ratio of the number of effective breeders to the census number of breeders would be 0.5. Mating between males and females was random and polygamous.

Standard demographic theory tells us that the long-term growth or decline rate of such a population is given by the dominant eigenvalue of **A**. We refer to this dominant eigenvalue as $\boldsymbol{h}(w)$ to emphasize that it depends on w. We impose density dependence in our model by setting w so that each year $\boldsymbol{h}(w) = 1 + (g - 1)(1 - \frac{N}{N_K})$, where $N_K = 1000$ is the carrying capacity and N is the total number of individuals in the population between the ages of one and eight, inclusive. The parameter g, determines the intrinsic rate of growth of the population. It is similar to r in equation 5, but it applies to growth of an age-structured population. We fixed the value of g to be 1.4. Values of w were highest in the first years and dropped off after that. The largest value of w, 1.5, occurred during the first few years for $c/2 = 20$. This corresponds to three-year-old females producing on average three offspring that survive to age one, and eight-year-old females producing 18 offspring that survive to age one. These are fairly high growth rates considering the low fecundity of resident *O. mykiss*, and the high expected mortality in the first year of life. Example trajectories of simulated populations are shown in Figure 8.

We combined our Monte Carlo estimates of $P(A(T)|n, N(t))$ with the values of $L(A(T) = a|\boldsymbol{x}_0, \boldsymbol{y}_0)$ using (10) to obtain values of $L(N(t))$, which can be regarded as a likelihood for the founding number of chromosomes, conditional on our growth model for the above-falls population. The MLE of the number of founding individuals is 421. Figure 9 shows that the log-likelihood curve rises steeply up to 120 founding individuals, then begins to slowly level off and drop back down. The two-unit support limit puts the lower endpoint of a one-sided 97.5% confidence interval at 185. This result suggests that, even with the generous assumed growth rate we applied to these populations, the number of juvenile fish transferred above the falls in 1906 would have to be greater than 185 to result in the genetic patterns observed today.

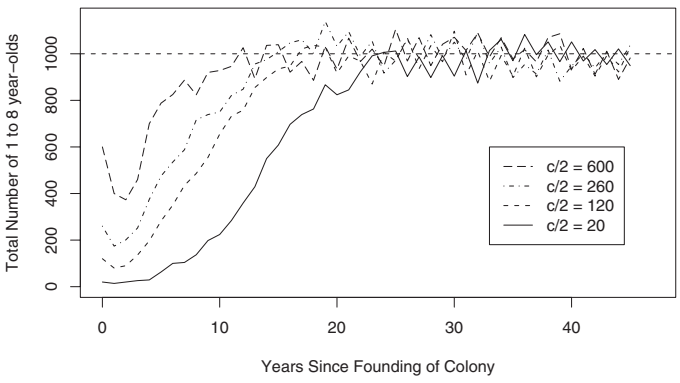

Figure 8. Simulated population sizes. Each curve shows the total population size (age one through eight) corresponding to one realization of the *spip* simulation. The different curves correspond to different numbers of founding individuals, as shown in the legend.

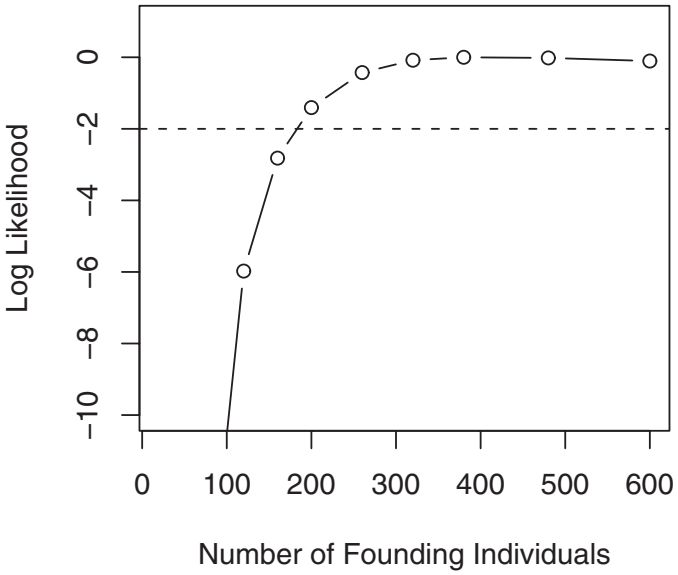

Figure 9. Log-likelihood curve of number of founding individuals ($c/2$) for the *O. mykiss* dataset.

Discussion and Conclusions

We have shown that it is possible to estimate the number of founding lineages ancestral to a sample of genes from a founded colony given genetic data at a locus from the colony and from its source. We have also shown that under suitably restricted conditions, it is possible to estimate the actual number of individuals (or chromosomes) present amongst the colonizers. Only when population growth is exceptionally rapid and the number of founders very small is the number of founding lineages close to the number of founding chromosomes, as was noted by Knowles et al. (1999) for mitochondrial DNA. In general, not all founding genes will leave descendents, even though it is likely that all or nearly all founding chromosomes will leave descendant lineages at some loci. Consequently, estimating c, the number of founding chromosomes, requires either that the growth rate of the founded colony is very high and c is low, or that a model, such as the logistic model, and a growth rate, may be assumed for the growth of the colony. Although the growth rate may not be known with accuracy, it is still possible to bound the likely range of c given assumed values of the population growth rate. This approach is clearly limited by the

fact that many different models for and rates of population growth are possible, and yet only one is being assumed and conditioned upon for the analysis. Even species with the potential for high growth rates (like some fishes or insects) may experience population growth after colony establishment characterized by extreme fluctuations. If the population size fluctuates down to or below the number of founders, then there is no method we know of that could accurately estimate the number of founders. However, even in such a difficult scenario, the logistic model can provide a reasonable estimate of the minimum number of founders. This was done for the *O. mykiss* dataset: even though the population was granted a generously high growth rate and asymptotically monotonic logistic growth, it was still apparent that the number of founding individuals would have to be quite large (>185) for the genetic data to be consistent with the hypothesis that the above-falls fish were derived exclusively from the 1906 transplants. If the population fluctuated wildly, then the number of founding individuals would have to be even higher.

The method described in this paper is applicable to loci with several distinguishable alleles, including allozyme and microsatellite loci. It is closely related to a method developed by Leblois and Slatkin (2007), which is applicable to closely linked single nucleotide polymorphisms (SNPs).

To provide an efficient calculation of the likelihood, we chose to assume that the genetic drift in the source population is negligible. This does not seem to bias the MLE of c when the true amount of drift is low; however, it may lead to overestimated precision of the MLE for c. We tried adopting several approximations to more adequately represent the increased uncertainty due to genetic drift in the source, but none were successful (results not shown). An additional approximation in the model is the assumption that Hoppe's urn without mutation (eq. 7) faithfully represents the neutral coalescent forward in time in the colony. If c is small, then the ancestry of a gene is likely to include coalescent events in which three or more lineages coalesce into one. Such events violate the assumptions that give rise to (7), but it is apparent from our results that the occurrence of multilineage coalescent events do not affect inference of c appreciably.

Our statistical method uses the coalescent process and explicitly includes and calculates $L(A(T) = a|\boldsymbol{x}_0, \boldsymbol{y}_0)$, the likelihood that the sample from the colony descended from a ancestral lineages extant at time T, given the genetic data. As with many calculations involving the coalescent conditioned on data, computing this quantity is difficult; however, approximating it using the importance sampling algorithm of Anderson (2005) can be done quickly. To simply estimate c accurately, it requires about 250 importance sampling replicates per value of a at each locus. For the trout dataset, this required 30 sec on a 2 GHz G5 processor. Obtaining accurate estimates of $L(A(T) = a|\boldsymbol{x}_0, \boldsymbol{y}_0)$ requires more importance sampling replicates. The curves in Figure 7 were obtained using 100,000 importance sampling replicates that required 3.3 h on the same processor. The software *nfcone* for performing these calculations is available for free download from http://santacruz.nmfs.noaa.gov/staff/eric_anderson/.

The quantity $L(A(T) = a|\boldsymbol{x}_0, \boldsymbol{y}_0)$ arises in other genetic inference problems when they are viewed from the coalescent perspective. It arises, for instance in Beaumont's (2003) method for estimating population growth or decline over time. A more elaborate version, which includes the possibility of mutation, arises in single-sample estimators of growth rate and effective population size (Kuhner et al. 1998). The importance sampling scheme used in *nfcone* might provide a novel way of generating proposal distributions in the Markov chain Monte Carlo algorithms required to compute the likelihood in such models, and could prove useful in extending such models to allow for samples taken at different times. Notably *nfcone* could be adapted to provide a test for loci under selection (or linked to loci under selection) caused by shifts in ecology or the invasion of novel habitats (Orr and Smith 1998).

ACKNOWLEDGMENTS

This research was supported in part by a grant from the National Institutes of Health (R01-GM40282) to MS. We thank D. Pearse of the Southwest Fisheries Science Center for sharing his genetic data from *O. mykiss* collected by S. Hayes and others, and we thank S. Hayes of the SWFSC for assistance in parameterizing the population model for the above-falls trout population. We are grateful to two anonymous referees who read the manuscript closely and provided helpful comments. The idea for this project arose in discussions with N. Ferrand.

LITERATURE CITED

Anderson, E. C. 2005. An efficient Monte Carlo method for estimating N_e from temporally spaced samples using a coalescent-based likelihood. Genetics 170:955–967.

Anderson, E. C., and K. K. Dunham. 2005. spip 1.0: a program for simulating pedigrees and genetic data in age-structured populations. Mol. Ecol. Notes 5:459–461.

Beaumont, M. 1999. Detecting population expansion and decline using microsatellites. Genetics 153:2013–2029.

———. 2003. Estimation of population growth or decline in genetically monitored populations. Genetics 164:1139–1160.

Carson, H., and A. Templeton. 1984. Genetic revolutions in relation to speciation phenomena: the founding of new populations. Annu. Rev. Ecol. Syst. 15:97–131.

Chikhi, L., M. W. Bruford, and M. A. Beaumont. 2001. Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. Genetics 158:1347–1362.

Edwards, A. W. F. 1992. Likelihood. Johns Hopkins Univ. Press, Baltimore, MD.

Griffiths, R. C., and S. Tavaré. 1994. Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. Lond. B Biol. Sci. 344:403–410.

Hoppe, F. 1984. Polya-like urns and the Ewen's sampling formula. J. Math. Biol. 20:91–94.

Kingman, J. F. C. 1982. On the genealogy of large populations. J. Appl. Prob. 19A:27–43.

Knowles, L. L., D. J. Futuyma, W. F. Eanes, and B. Rannala. 1999. Insight into speciation from historical demography in the phytophagous beetle genus *Ophraella*. Evolution 53:1846–1854.

Kuhner, M. K., J. Yamato, and J. Felsenstein. 1998. Maximum likelihood estimation of population growth rates based on the coalescent. Genetics 149:429–434.

Luikart, G., F. W. Allendorf, J. M. Cornuet, and W. B. Sherwin. 1998a. Distortion of allele frequency distributions provides a test for recent population bottlenecks. J. Hered. 89:238–247.

Luikart, G., J. M. Cornuet, and F. W. Allendorf. 1999. Temporal changes in allele frequencies provide estimates of population bottleneck size. Conserv. Biol. 13:523–530.

Luikart, G., W. B. Sherwin, B. M. Steele, and F. W. Allendorf. 1998b. Usefulness of molecular markers for detecting population bottlenecks via monitoring genetic change. Mol. Ecol. 7:963–974.

Mayr, E. 1954. Change of genetic environment and evolution. Pp. 157–180 *in* J. Huxley, A. C. Hardy, and E. B. Ford, eds., Evolution as a process. Allen and Unwin, London.

Nei, M., T. Maruyama, and R. Chakraborty. 1975. The bottleneck effect and genetic variability in populations. Evolution 29:1–10.

Orr, M. R., and T. B. Smith. 1998. Ecology and speciation. Trends Ecol. Evol. 13:502–506.

Risch, N., H. Tang, H. Katzenstein, and J. Ekstein. 2003. Geographic distribution of disease mutations in the Ashkenazi Jewish population supports genetic drift over selection. Am. J. Hum. Genet. 72:812–22.

Slatkin, M. 2004. A population-genetic test of founder effects and implications for ashkenazi jewish diseases. Am. J. Hum. Genet. 75:282–293.

Tavaré, S. 1984. Lines of descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. 26:119–164.

Vogel, F., and A. G. Motulsky. 1996. Human genetics: problems and approaches. 3rd ed.Springer, New York.

Wright, S. 1931. Evolution in Mendelian populations. Genetics 16:97–159.

———. 1937. The distribution of gene frequencies in populations. Proc. Natl. Acad. Sci. U.S.A. 23:307–320.

———. 1938. Size of population and breeding structure in relation to evolution. Science (Wash. D.C.) 87:430–431.

Associate Editor: C. Goodnight