Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences

A. J. Drummond, A. Rambaut, B. Shapiro, and O. G. Pybus

Department of Zoology, University of Oxford, Oxford, United Kingdom

We introduce the Bayesian skyline plot, a new method for estimating past population dynamics through time from a sample of molecular sequences without dependence on a prespecified parametric model of demographic history. We describe a Markov chain Monte Carlo sampling procedure that efficiently samples a variant of the generalized skyline plot, given sequence data, and combines these plots to generate a posterior distribution of effective population size through time. We apply the Bayesian skyline plot to simulated data sets and show that it correctly reconstructs demographic history under canonical scenarios. Finally, we compare the Bayesian skyline plot model to previous coalescent approaches by analyzing two real data sets (hepatitis C virus in Egypt and mitochondrial DNA of Beringian bison) that have been previously investigated using alternative coalescent methods. In the bison analysis, we detect a severe but previously unrecognized bottleneck, estimated to have occurred 10,000 radiocarbon years ago, which coincides with both the earliest undisputed record of large numbers of humans in Alaska and the megafaunal extinctions in North America at the beginning of the Holocene.

Introduction

Considerable progress in the field of population genetic inference has been made during the past decade, following parallel increases in computer processing speed and available gene sequence data. Most current methods are based on coalescent theory, a stochastic process that describes how population genetic processes determine the shape of the genealogy of sampled gene sequences. Coalescent-based inference methods enable population genetic parameters to be estimated directly from gene sequence data under a variety of scenarios, including recombination (Griffiths and Marjoram 1996; Kuhner, Yamato, and Felsenstein 2000; Fearnhead and Donnelly 2001), population subdivision (Bahlo and Griffiths 2000; Beerli and Felsenstein 2001), and variable population size (Kuhner, Yamato, and Felsenstein 1998; Beaumont 1999; Drummond et al. 2002).

The variable population size coalescent model, introduced by Griffiths and Tavare (1994) and Donnelly and Tavare (1995), allows the inference of past population dynamics from contemporary gene sequences. The model was subsequently extended by Rodrigo and Felsenstein (1999) to sequences that have been sampled at significantly different points in time (e.g., rapidly evolving pathogen sequences or ancient DNA sequences). The inference of demographic history from genetic data has proved invaluable for testing hypotheses in a variety of biological disciplines, for example, anthropology (Reich and Goldstein 1998), epidemiology (Pybus et al. 2001; Joy et al. 2003), conservation biology (Roman and Palumbi 2003), and ecology (Storz, Beaumont, and Alberts 2002; Flanagan et al. 2004).

Coalescent methods for inferring demographic histories require a "demographic model," which is simply a mathematical function used to describe the change in effective population size through time. Each demographic model has one or more "demographic parameters." Commonly used demographic models are constant size (one parameter), exponential growth (constant growth rate through time; two parameters), logistic growth (decreasing growth rate through time; three parameters), and expansion growth (increasing growth rate through time; three parameters) (see Pybus and Rambaut [2002] for more details). In addition, combining the models above in a piecewise manner enables an array of more complex models to be constructed. Past population dynamics are reconstructed by estimating the demographic parameters, typically by maximum likelihood or Bayesian methods (Kuhner, Yamato, and Felsenstein 1998; Pybus, Rambaut, and Harvey 2000; Drummond et al. 2002).

It is not usually known in advance which demographic model will fit the gene sequences under investigation. Although it is possible to compare the fit of different demographic models using standard model selection techniques (Pybus, Rambaut, and Harvey 2000), this is a time-consuming process and there is no guarantee that any of the compared models will fit the data adequately. The use of an incorrect demographic model will lead to biased and invalid estimates of demographic history. Furthermore, an incorrect demographic model could lead to bias in other estimated parameters of the evolutionary model, such as recombination rate or overall mutation rate, when the demographic history is being treated as a nuisance parameter to be averaged over. To address the first of these concerns, a flexible model called the skyline plot was developed (Pybus, Rambaut, and Harvey 2000). The "skyline plot" is a piecewise-constant model of population size that can fit a wide range of demographic scenarios. It has proved very useful as a model selection tool that is used to indicate the most appropriate demographic model for any given data set (Pybus and Rambaut 2002). The skyline plot typically produces "noisy" plots that display the stochastic variability inherent in the coalescent process. To reduce this noise, the "generalized skyline plot" was developed (Strimmer and Pybus 2001), which uses the Akaike Information Criterion (Akaike 1974) to reduce the number of parameters employed and thus produces smoother estimated population size plots.

Like early genealogical estimators of population size (Felsenstein 1992; Fu 1994), both of the skyline plot methods have a fundamental drawback—they infer demographic history from an estimated genealogy, rather than from the sampled gene sequences, and thus ignore the error associated

Key words: coalescent inference, Markov chain Monte Carlo, demographic model selection, skyline plot, megafaunal extinctions, Holocene.

E-mail: alexei.drummond@zoo.ox.ac.uk.

Mol. Biol. Evol. 22(5):1185–1192. 2005
doi:10.1093/molbev/msi103
Advance Access publication February 9, 2005

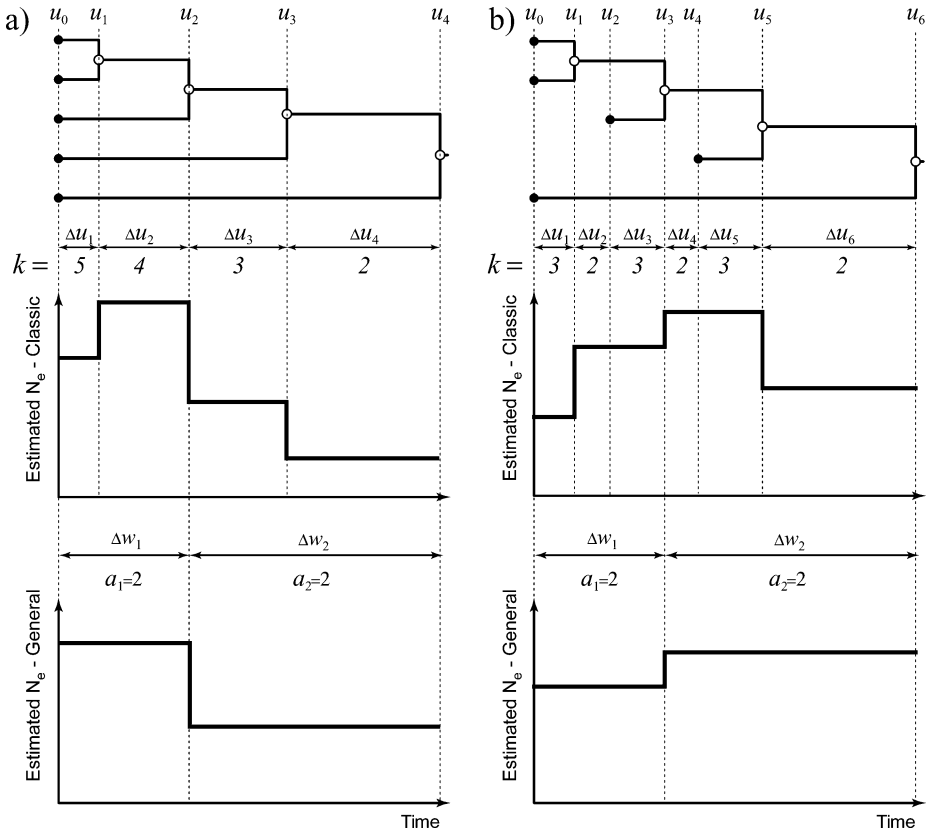

FIG. 1.—(*a*) A genealogy of five individuals sampled contemporaneously (top) together with its associated classic (middle) and generalized (bottom) skyline plots. (*b*) A genealogy of five individuals sampled at three different times (top) along with its associated classic (middle) and generalized (bottom) skyline plots. In the classic skyline plots, the changes in effective population size coincide with coalescent events, resulting in a stepwise function with $n - 2$ change points and $n - 1$ population sizes, where n is the number of sampled individuals. In the generalized skyline plot, changes in effective population size coincide with some, but not necessarily all, coalescent events. The resulting stepwise function has $m - 1$ change points $(1 \leq m \leq n - 1)$ and m population sizes.

with phylogenetic reconstruction. Although this error may be small for highly variable sequences, such as those from rapidly evolving viruses (Drummond, Pybus, and Rambaut 2003; Drummond et al. 2003), it is impractical to ignore phylogenetic error in the majority of less variable data sets. Here we resolve this problem by introducing the "Bayesian skyline plot." The Bayesian skyline plot model uses standard Markov chain Monte Carlo (MCMC) sampling procedures to estimate a posterior distribution of effective population size through time directly from a sample of gene sequences, given any specified nucleotide-substitution model. Unlike previous methods, the Bayesian skyline plot includes credibility intervals for the estimated effective population size at every point in time, back to the most recent common ancestor of the gene sequences. These credibility intervals represent both phylogenetic and coalescent uncertainty. In addition, the "averaging" effect of MCMC sampling naturally produces smoother estimates than previous skyline plots.

First, we generalize previous descriptions of the generalized skyline plot and show how it can be extended to trees relating sequences sampled at different points in time (termed heterochronous trees). Second, we describe how the generalized skyline plot model can be embedded in a Bayesian MCMC analysis. Third, we introduce an MCMC sampling procedure that efficiently samples the distribution of generalized skyline plots given the sequence data and combines these plots to generate a posterior distribution of effective population size through time. By doing this, we are able to coestimate the evolutionary rate, substitution model parameters, phylogeny, and ancestral population dynamics within a single analysis. Fourth, we apply the Bayesian skyline plot to simulated data sets and show that it correctly reconstructs demographic history under canonical scenarios. Last, we compare the Bayesian skyline plot model to previous approaches by analyzing two real data sets that have been investigated using other coalescent-based methods.

Methods

Background: Classic and Generalized Plots

The classic and generalized skyline plots were introduced by Pybus, Rambaut, and Harvey (2000) and Strimmer and Pybus (2001), respectively. The raw data for the classic and generalized skyline plots are a genealogy with specified branch lengths, denoted g, estimated from n contemporaneous sequences (see fig. 1*a*). The internal nodes of the input genealogy must be dated according to a given timescale (genetic distance or time) and thus define $n - 1$ times at which coalescent events occur, $\mathbf{u} = \{u_1, u_2, \ldots, u_{n-1}\}$

(see fig. 1*a*). The times **u** are measured from the tips, such that $u_0 = 0$ at the tips. The waiting times between coalescent events (coalescent intervals) are defined $\Delta u_i = u_i - u_{i-1}$. The number of lineages present during each Δu_i defines a corresponding series of values, denoted $k_1, k_2, \ldots, k_{n-1}$ (see fig. 1*a*).

Strimmer and Pybus (2001) demonstrated that the skyline plot is, in fact, a method-of-moments estimate of a piecewise model of effective population size through time and therefore has a definable likelihood function. The piecewise model supposes that effective population size is constant between coalescent events but may change at the coalescent event times, **u**. The classic skyline plot assumes that each coalescent interval has a different effective population size, whereas the generalized skyline plot allows adjacent coalescent intervals to be grouped and has the same effective population size (fig. 1*a*).

The generalized skyline plot therefore requires an ordered subset of group sizes $A=\{a_1, a_2, \ldots, a_m\}$ ($a_i > 0$ and $\sum_{i=1}^{m} a_i = n-1$) that defines the number of coalescent events in each grouped interval. Here m is the number of grouped intervals ($1 \leq m \leq n-1$). The time at which each grouped interval ends is denoted $\mathrm{w}=\{w_1, w_2, \ldots, w_m\}$ and is a subset of **u** ($w_0 = 0$). The time spanned by each grouped interval is defined $\Delta w_j = w_j - w_{j-1}$ (fig. 1*a*).

The log likelihood of the piecewise demographic model is given by:

$$\log f_G(g \mid \Theta, A) = \sum_{i=1}^{n-1} \log\frac{k_i(k_i-1)}{2\theta_{h(i)}} - \frac{k_i(k_i-1)\Delta u_i}{2\theta_{h(i)}}. \quad (1)$$

The vector $\Theta=\{\theta_1, \theta_2, \ldots, \theta_m\}$ is the vector of effective population sizes for each of the grouped intervals Δw_j. The function $h(\,)$ provides a mapping from the indices of **u** to the indices of **w** and is defined:

$$h(i) := \begin{cases} 1, & \text{if } i \leq a_1, \\ j, & \text{if } \sum_{k=1}^{j-1} a_k < i \leq \sum_{k=1}^{j} a_k. \end{cases} \quad (2)$$

Likelihoods of Skyline Plots for Heterochronous Trees

The generalized skyline plot approach defined above is only applicable to gene sequences sampled at the same point in time. Here we show how the skyline plot approach can be extended to heterochronous sequences, using the serial-sample coalescent model (Rodrigo and Felsenstein 1999).

Figure 1*b* shows a genealogy of heterochronous sequences with internal nodes and sampling times dated according to a given timescale. In contrast to a genealogy of contemporaneous sequences, heterochronous genealogies have two types of intervals: coalescent intervals (which, going back in time, end with a coalescent event) and sample intervals (which end with the appearance of one or more sampled sequences; a sample event). If there are n sequences sampled at s different times, then there are $n+s-2$ intervals in total (fig. 1*b*). The ordered interval times, starting at the tips and ending at the root, are denoted $u_1, u_2, \ldots, u_{n+s-2}$, and an indicator function, $I_c(i)$, is used to indicate whether the ith event is a coalescent event. $I_c(i)=1$ if the event is a coalescent event, otherwise $I_c(i)=0$. The values of Δu_i and k_i are now associated with either sample-ended or coalescent-ended intervals (see fig. 1*b*). Note also that the number of lineages, k_i, can both increase and decrease going back in time because sample events add lineages and coalescent events remove them.

Coalescent-ended intervals in heterochronous trees can be grouped in a manner similar to that described above. As before, the values $A=\{a_1, a_2, \ldots, a_m\}$ define the number of coalescent events in each grouped interval. The time at which each grouped interval ends is denoted $\mathbf{w} = \{w_1, w_2, \ldots, w_m\}$, representing a subset of **u**, where m is the number of grouped intervals $(1 \leq m \leq n-1)$and $w_0=0$. Note that the group times, **w**, only contain coalescent events, not sample events. The time spanned by each grouped interval is defined $\Delta w_j = w_j - w_{j-1}$ (fig. 1*b*).

As before, $\Theta=\{\theta_1, \theta_2, \ldots, \theta_m\}$ represents the effective population size within each grouped interval. Effective population size does not change at the end of sample-ended intervals. Thus, the log likelihood of the piecewise demographic model is given by:

$$\log f_G(g \mid \Theta, A) = \sum_{i=1}^{n+s-2} I_c(i)\log\frac{k_i(k_i-1)}{2\theta_{h(i)}} - \frac{k_i(k_i-1)\Delta u_i}{2\theta_{h(i)}}, \quad (3)$$

where the function $h(\,)$ again provides a mapping from indices in **u** to indices in **w** and is defined:

$$h(i) := \begin{cases} 1, & \text{if } \sum_{j=1}^{i} I_c(j) \leq a_1, \\ j, & \text{if } \sum_{k=1}^{j-1} a_k < \sum_{j=1}^{i} I_c(j) \leq \sum_{k=1}^{j} a_k. \end{cases} \quad (4)$$

The Bayesian Skyline Plot Model

The vectors $\Theta=\{\theta_1, \theta_2, \ldots, \theta_m\}$ and $A=\{a_1, a_2, \ldots, a_m\}$, together with a genealogy (g), define a piecewise demographic history with $2m-1$ demographic parameters and $n-1$ coalescent time parameters. The probability $f_G(g \mid \Theta, A)$ of the genealogy (g) given the demographic parameters is given in equation (3). The hyperparameter m is either chosen a priori or is sampled in a hierarchical model via reversible-jump MCMC (rjMCMC).

In addition to this model, we also introduce a simple smoothing on Θ which represents our belief that effective population size is autocorrelated through time. The prior distribution we assume in all subsequent simulations and analyses is that, going back in time, each new population size is drawn from an exponential distribution with a mean equal to the previous population size:

$$\theta_j \sim \mathrm{Exp}(\theta_{j-1}), \quad 2 \leq j \leq m. \quad (5)$$

In addition, we introduce a scale-invariant prior (Jeffreys 1946) on the first element $f_{\theta_1}(\theta_1) \propto 1/\theta_1$ to signify that

our prior belief is invariant to changes in timescale. This results in the following multivariate prior distribution on Θ:

$$f_{\Theta}(\Theta) \propto \frac{1}{\theta_1} \prod_{j=2}^{m} \theta_{j-1} \exp(-\theta_j/\theta_{j-1}). \quad (6)$$

MCMC Implementation

Here we extend a previously published MCMC method (Drummond et al. 2002) to sample the parameters of the Bayesian skyline plot model described above. Our implementation samples both Θ and A, but for computational reasons, we do not use rjMCMC to sample the hyperparameter m. Instead, we condition all our runs on a fixed value of m because the resulting posterior demographic function is highly consistent for a range of a priori values of m (data not shown).

In the case of contemporaneous sequences, the posterior distribution sampled in our scheme is:

$$f_{iso}(\Theta, A, \Omega, g \mid D, \mu) = \frac{1}{Z} \Pr\{D \mid \mu, g\} f_G(g \mid \Theta, A) \times f_{\Theta}(\Theta) f_A(A) f_{\Omega}(\Omega). \quad (7)$$

In equation (7), the vector Ω contains the parameters of the substitution model (such as transition/transversion ratio [κ], shape parameter for gamma-distributed rate across sites [α], and proportion of invariant sites [p_{inv}]), and the parameter μ is a mutation rate that scales the genealogy from units of mutations per site to units of time. In the case of heterochronous sequences, the posterior distribution sampled in our scheme is:

$$f_{het}(\Theta, A, \Omega, g, \mu \mid D) = \frac{1}{Z} \Pr\{D \mid \mu, g\} f_G(g \mid \Theta, A) f_{\Theta}(\Theta) \times f_A(A) f_{\Omega}(\Omega) f_{\mu}(\mu). \quad (8)$$

It should be noted that if $f_{\mu}(\mu)$ is chosen to be a probability density function (i.e., a proper prior), then this second scheme could also be used for contemporaneous sequences to allow uncertainty in the scaling between mutations per site and time.

The resulting output of the Bayesian MCMC analysis is a sample from the posterior distribution described in equation (8), $(\Theta, A, \Omega, g, \mu) \sim f_{het}$. For the purposes of reconstructing the demographic history, the sampled substitution model parameters and mutation rates can be thought of as uninteresting "missing data." This leaves us with a list of j states, each with an associated genealogy and demographic parameters: (Θ_j, A_j, g_j). We can therefore reconstitute the demographic history $\theta(t)$ as a piecewise function of time, for each of the j states. To display this information, we calculate the marginal posterior distribution of $\theta(t)$ at a series of times of interest, t_i. For each t_i, the values $\theta(t_i)$ for all j form a marginal posterior distribution of population size at time t_i. From these marginal distributions, the mean, median, and the 95% highest posterior density (HPD) intervals can be calculated for $\theta(t)$ at each time of interest, t_i, leading to an estimated plot of population size through time with an associated measure of uncertainty.

Results

Simulated Data Sets

To investigate the behavior of the Bayesian skyline plot model, we analyzed two simulated data sets using our MCMC method. We followed Strimmer and Pybus (2001) and performed the following simulations: (1) Coalescent trees were simulated under two demographic models, $\theta(t) = 0.05$ (constant) and $\theta(t) = e^{-1000t}$ (exponential). These models approximately represent the history and genetic diversity of animal mitochondrial DNA (mtDNA) sequences. In these demographic models, time is measured in mutations per site. (2) Sequences were simulated down the trees using the HKY (Hasegawa, Kishino, and Yano 1985) model (transition/transversion ratio = 10; nucleotide frequencies A = 0.3, C = 0.25, G = 0.15, and T = 0.3) with a uniform mutation rate among sites. The constant-model sequence alignment was 500 base pairs (bp) in length, and the exponential-model alignment was 1,500 bp long. (3) The resulting sequence alignments were then analyzed using the MCMC method described above. In both analyses, the number of groups (m) was set to 12, and MCMC chains were run for 10,000,000 iterations, of which the first 1% was discarded to allow for burn-in. The substitution model used was HKY, and the transition/transversion ratio was coestimated along with the parameters of the Bayesian skyline plot and the ancestral genealogy. Genealogies and model parameters were sampled every 1,000 iterations, and the results are summarized as a Bayesian skyline plot, shown in figure 2.

As noted by Strimmer and Pybus (2001), many of the sequences simulated under the constant-size model are not unique and many of the coalescent intervals in the estimated tree are very small. The effect of this is most apparent at the tipward (most recent) end of the Bayesian skyline plot (fig. 2*a*) where the variance in the estimate of population size is larger than at other times in the plot. Under the exponential model, the simulation resulted in only two identical sequences. The Bayesian skyline plot appears to slightly overestimate population size in this data set; however, this is well within the uncertainty admitted by the 95% HPD limits (fig. 2*b*) and may well be due to the stochastic error associated with this particular simulation.

To compare the Bayesian skyline plot with previous approaches, we also analyzed two published data sets that have been investigated previously using other coalescent inference methods.

Hepatitis C Virus in Egypt

The hepatitis C virus (HCV) is a genetically diverse RNA virus and is a leading global cause of liver disease. Egypt has the highest prevalence of HCV in the world, significantly higher than those in neighboring countries, even though HCV appears to have been present in the Middle East for several centuries (Smith et al. 1997). The past dynamics of HCV in Egypt are therefore of considerable interest. Here we analyze a data set of 63 partial E1 gene sequences obtained from a comprehensive study of Egyptian HCV genetic diversity (Ray et al. 2000). The demographic history of these data was analyzed by Pybus et al. (2003)

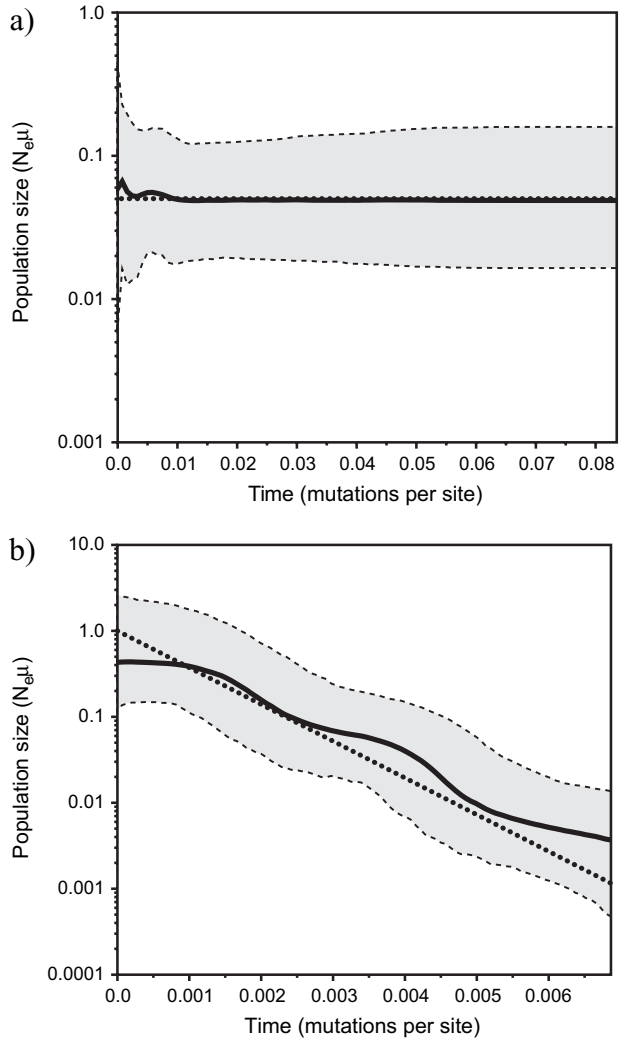

FIG. 2.—Performance of the Bayesian skyline plot on simulated data. Time is measured in units of mutations per site. The true demographic histories are shown as thick dotted lines, the median estimates are shown as thick solid lines, and the 95% HPD limits are shown by the gray areas bounded by thin dashed lines. (*a*) The Bayesian skyline plot ($m = 12$) calculated from a set of sequences that were simulated under a model of constant population size. (*b*) The Bayesian skyline plot ($m = 12$) calculated from a simulated data set for which the true demographic history was exponential growth (see text for details).

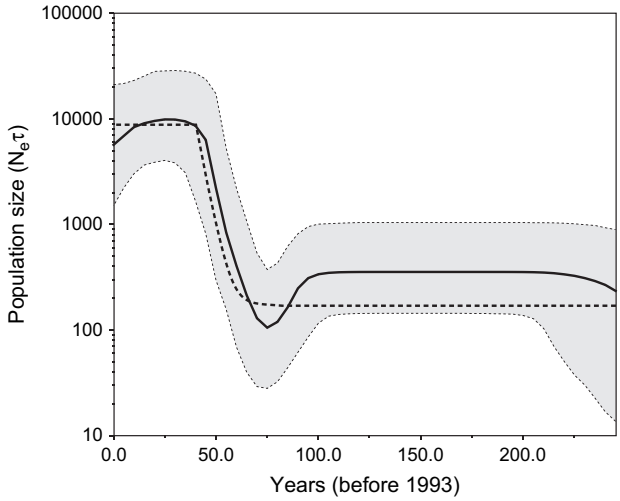

FIG. 3.—A Bayesian skyline plot ($m = 24$) derived from an alignment of Egyptian HCV sequences (63 partial E1 gene sequences, sampled in 1993). The *x* axis is in units of years before 1993, and the *y* axis is equal to $N_e\tau$ (the product of the effective population size and the generation length in years). The thick solid line is the median estimate, and the dashed lines show the 95% HPD limits. The thick dashed line shows the mean estimate for the four-parameter model used in Pybus et al. (2003) (see text for details). The plot shows a sharp increase in the effective number of infections in the early 20th century, probably caused by viral contamination of injectable antischistosomiasis treatment that was widely used in Egypt from 1920s (see text for details).

using a four-parameter coalescent model, estimated within a Bayesian inference framework. As previously noted, these sequences are suitable for coalescent analysis because (1) they represent a geographically diverse and approximately random sample, (2) they show no obvious population subdivision, (3) they contain ample phylogenetic information, and (4) an independent estimate of nucleotide-substitution rate for the sequences is available (see Pybus et al. [2003] for details). Most importantly, there is substantial nongenetic evidence concerning the history of HCV spread in Egypt (Frank et al. 2000; Strickland et al. 2002), so this data set provides a unique opportunity to test the reliability and accuracy of coalescent demographic methods.

The analysis of Pybus et al. (2003) demonstrated that the effective number of HCV infections underwent a dramatic period of growth in the middle of the 20th century, followed by a more recent slowdown in the rate of spread. Figure 3 shows the results of the Bayesian skyline plot ($m = 24$) on this data set, the HPD limits of which entirely contain the previously published estimate. The Bayesian skyline plot is also in complete agreement with known epidemiological data—the Egyptian HCV epidemic was likely caused by viral contamination of injectable antischistosomiasis treatment. This treatment was extensively used in Egypt from the 1920s to the 1980s but was phased out gradually during the latter part of this period. The Bayesian skyline plot shows a period of rapid HCV population increase between 1920 and 1950. The growth phase is preceded by a dip in the effective number of infections. This dip was not observed in the previous coalescent analysis but is not statistically significant given the size of the estimated confidence limits.

Bison in Beringia

Bison were one of the most abundant and widely distributed large mammals during the Late Pleistocene (ca. 500–10,000 years ago [1000 years ago, ka B.P.]). A recent coalescent analysis (Shapiro et al. 2004) of mtDNA control region sequences from ancient bison (*Bison* cf. *priscus*) in Beringia (Siberia, Alaska, and northwestern Canada) and central North America revealed a demographic history over the last 150 ka B.P. of sustained population growth followed by rapid decline. The estimated time of the transition from growth to decline (mean: 37 ka B.P.) led the authors to conclude that this change in demographic trend was more likely

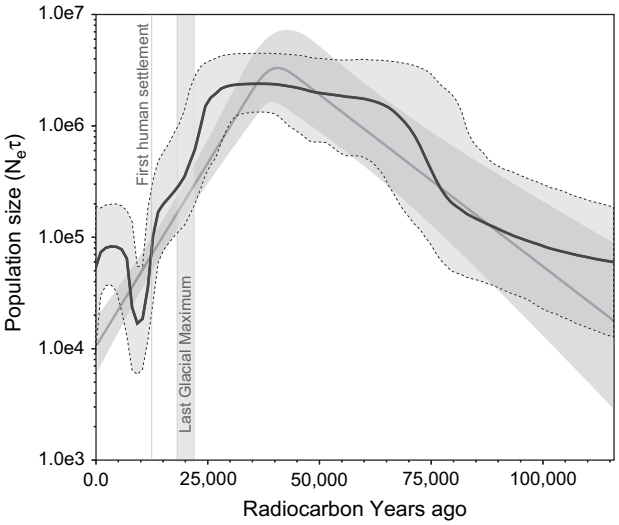

FIG. 4.—A Bayesian skyline plot ($m = 15$) derived from a sample ($n = 191$) of mtDNA control region sequences from ancient bison (*Bison* cf. *priscus*) preserved in permafrost in Beringia. The x axis is in units of radiocarbon years in the past, and the y axis is equal to $N_e\tau$ (the product of the effective population size and the generation length in radiocarbon years). The thick solid line is the median estimate, and the dashed lines show the 95% HPD limits. The darker underlay is the estimated median and 95% HPD limits of the four-parameter model used in Shapiro et al. (2004) (see text for details). The four-parameter model captures a large part of the underlying demographic signal; however, significant deviation between the two plots is evident in the most recent ~15,000 radiocarbon years. The Bayesian skyline plot suggests that the bison species went through a severe bottleneck around 10,000 radiocarbon years ago, coincident with the time when many North American megafaunal species went extinct. For comparison, the time of first human settlement in Alaska and the cold period around the Last Glacial Maximum are also indicated.

the consequence of climate changes associated with the onset of the Last Glacial Maximum than human overhunting. The authors used a four-parameter, two-epoch model to describe this demographic history, in which an initial phase of exponential growth was followed by a period of exponential decline. We reanalyzed the Shapiro et al. (2004) data using a Bayesian skyline plot with 15 groups ($m = 15$). The analysis was run for 30 million iterations, with the first 10% discarded as burn-in. Genealogies and model parameters were sampled every 1,000 iterations thereafter. Despite the increased number of parameters, a comparison of log marginal posterior scores between the two analyses shows that the Bayesian skyline plot is a significantly better fit to the data than the simpler model (data not shown).

Figure 4 shows the Bayesian skyline plot for Beringian bison, along with the demographic history estimated in Shapiro et al. (2004). Although the two analyses are very similar, the sensitivity of the Bayesian skyline plot allows more complex demographic trends to be identified. Most striking is the population bottleneck following the rapid decline of bison population size, between its maximum 30–45 ka B.P. and its minimum ~10 ka B.P. The postbottleneck recovery, which the simpler four-parameter model did not identify, is seen in the Bayesian skyline plot as a second transition to population growth around 10 ka B.P. This bottleneck is coincident with the earliest undisputed evidence of human settlement in Alaska (Yesner 2001), the period most often associated with the megafaunal mass extinctions in North America (Alroy 1999). In agreement with the Shapiro et al. (2004) analysis, the Bayesian skyline plot describes the initial transition from growth to decline as occurring prior to the first evidence of humans in North America; however, figure 4 suggests that humans could have played a larger role in the decline of bison populations than previously suggested.

Conclusions

The Bayesian skyline plot is a powerful new method for estimating ancestral population dynamics from a sample of molecular sequences. In common with the classic and generalized skyline plots, the Bayesian skyline plot allows us to discover novel demographic signatures that are not readily described by simple demographic models. Unlike previous skyline plots, our new method takes into account both the error inherent in phylogenetic reconstruction and the stochastic error intrinsic to the coalescent process and thus produces more correct estimates of statistical uncertainty. In addition, the MCMC method coestimates the ancestral genealogy and parameters of the substitution process as well as the demographic parameters.

In both the HCV and bison examples presented above, the Bayesian skyline plot reveals previously undetected demographic signatures, demonstrating its ability to uncover changes in demographic trends over ecological, paleontological, and evolutionary time spans. The Bayesian skyline plot may also prove useful when the aim of inference is a population genetic parameter such as recombination or mutation rate, and the demographic history of the sequences is a nuisance parameter that must be "averaged out." In such circumstances, an unrealistic or misspecified model of population size change may lead to strong biases in the estimated parameter of primary interest. Ascertaining the direction and magnitude of such a bias is outside the scope of this paper because it will depend on the data set and sampling strategy. However, given the nontrivial demographic histories estimated from real data sets in this paper, estimates of mutation rate or recombination rate obtained by assuming a simple demographic history should probably be supported with simulations that verify the robustness of the focal parameter estimates to the simplifying demographic assumptions.

The Bayesian skyline plot takes a similar amount of time to compute as other demographic models (such as constant size or exponential growth) estimated under the same MCMC genealogy sampling method (Drummond et al. 2002). For example, the HCV data set took a few hours to calculate on a desktop computer. However, this is substantially longer than the classic and generalized skyline plots estimators, which typically take a few seconds to compute (although obviously this does not include the time necessary to estimate the tree).

During preparation of the manuscript, we became aware of a related method (Opgen-Rhein, Fahrmeir, and Strimmer 2005) that uses rjMCMC to estimate smooth demographic functions directly from a single reconstructed

genealogy. This approach, unlike the Bayesian skyline plot, does not take phylogenetic error into account and is therefore less appropriate for data sets containing limited genetic variation. However, the use of rjMCMC by Strimmer and coworkers is preferable to our method of fixing the amount of smoothing a priori (i.e., fixing the number of groupings, m). The two methods are therefore complementary especially as they use different information as starting data.

Future improvements to the Bayesian skyline plot method should further extend its utility. Future research directions include (1) implementation of an rjMCMC method so that the degree of grouping can be automatically chosen from the data, (2) use of a smoother function, such as piecewise linear or truncated regression spline, as the underlying parametric model, and (3) extension of the Bayesian skyline plot to sequence data from multiple unlinked loci.

The MCMC method described in this paper is implemented in the latest version of the BEAST (v. 1.2) software package (available from http://evolve.zoo.ox.ac.uk/beast/). The Bayesian skyline plot figures presented in this paper were generated using Tracer (v. 1.2) (available from http://evolve.zoo.ox.ac.uk/software/).

Acknowledgments

We are grateful to Korbinian Strimmer for providing us early access to a stimulating manuscript describing related research. In addition, A.J.D. thanks Chris Holmes and Alexandre Pintore for helpful discussions. A.J.D. and B.S. are funded by the Wellcome Trust. A.R. and O.G.P. are funded by the Royal Society.

Literature Cited

Akaike, H. 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control **AC-19**:716–723.

Alroy, J. 1999. Putting North America's end-Pleistocene megafaunal extinction in context: large-scale analyses of spatial patterns, extinction rates, and size distributions. Pp. 105–143 *in* R. D. E. MacPhee, ed. Extinctions in near time: causes, contexts, and consequences. Kluwer Academic, New York.

Bahlo, M., and R. C. Griffiths. 2000. Inference from gene trees in a subdivided population. Theor. Popul. Biol. **57**:79–95.

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

Beerli, P., and J. Felsenstein. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc. Natl. Acad. Sci. USA **98**:4563–4568.

Donnelly, P., and S. Tavare. 1995. Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. **29**:401–421.

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

Drummond, A., O. G. Pybus, and A. Rambaut. 2003. Inference of viral evolutionary rates from molecular sequences. Adv. Parasitol. **54**:331–358.

Drummond, A. J., O. G. Pybus, A. Rambaut, R. Forsberg, and A. G. Rodrigo. 2003. Measurably evolving populations. Trends Ecol. Evol. **18**:481–488.

Fearnhead, P., and P. Donnelly. 2001. Estimating recombination rates from population genetic data. Genetics **159**:1299–1318.

Felsenstein, J. 1992. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. **59**:139–147.

Flanagan, N. S., A. Tobler, A. Davison, O. G. Pybus, D. D. Kapan, S. Planas, M. Linares, D. Heckel, and W. O. McMillan. 2004. Historical demography of Mullerian mimicry in the Neotropical Heliconius butterflies. Proc. Natl. Acad. Sci. USA **101**: 9704–9709.

Frank, C., M. K. Mohamed, G. T. Strickland et al. (11 co-authors). 2000. The role of parenteral antischistosomal therapy in the spread of hepatitis C virus in Egypt. Lancet **355**:887–891.

Fu, Y. X. 1994. A phylogenetic estimator of effective population size or mutation rate. Genetics **136**:685–692.

Griffiths, R. C., and P. Marjoram. 1996. Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol. **3**:479–502.

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

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160–174.

Jeffreys, H. 1946. An invariant form for the prior probability in estimation problems. Proc. R. Soc. A **186**:453–461.

Joy, D. A., X. Feng, J. Mu et al. (12 co-authors). 2003. Early origin and recent expansion of Plasmodium falciparum. Science **300**:318–321.

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

———. 2000. Maximum likelihood estimation of recombination rates from population data. Genetics **156**:1393–1401.

Opgen-Rhein, R., L. Fahrmeir, and K. Strimmer. 2005. Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. BMC Evol. Biol. **5**:6.

Pybus, O. G., M. A. Charleston, S. Gupta, A. Rambaut, E. C. Holmes, and P. H. Harvey. 2001. The epidemic behavior of the hepatitis C virus. Science **292**:2323–2325.

Pybus, O. G., A. J. Drummond, T. Nakano, B. H. Robertson, and A. Rambaut. 2003. The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. Mol. Biol. Evol. **20**:381–387.

Pybus, O. G., and A. Rambaut. 2002. GENIE: estimating demographic history from molecular phylogenies. Bioinformatics **18**:1404–1405.

Pybus, O. G., A. Rambaut, and P. H. Harvey. 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics **155**:1429–1437.

Ray, S. C., R. R. Arthur, A. Carella, J. Bukh, and D. L. Thomas. 2000. Genetic epidemiology of hepatitis C virus throughout Egypt. J. Infect. Dis. **182**:698–707.

Reich, D. E., and D. B. Goldstein. 1998. Genetic evidence for a Paleolithic human population expansion in Africa. Proc. Natl. Acad. Sci. USA **95**:8119–8123.

Rodrigo, A. G., and J. Felsenstein. 1999. Coalescent approaches to HIV population genetics. Pp. 233–272 *in* K. Crandall, ed. Molecular evolution of HIV. Johns Hopkins University Press, Baltimore, Md.

Roman, J., and S. R. Palumbi. 2003. Whales before whaling in the North Atlantic. Science **301**:508–510.

Shapiro, B., A. J. Drummond, A. Rambaut et al. (27 co-authors). 2004. Rise and fall of the Beringian steppe bison. Science **306**:1561–1565.

Smith, D. B., S. Pathirana, F. Davidson, E. Lawlor, J. Power, P. L. Yap, and P. Simmonds. 1997. The origin of hepatitis C virus genotypes. J. Gen. Virol. **78**(Pt 2):321–328.

Storz, J. F., M. A. Beaumont, and S. C. Alberts. 2002. Genetic evidence for long-term population decline in a Savannah-dwelling primate: inferences from a hierarchical Bayesian model. Mol. Biol. Evol. **19**:1981–1990.

Strickland, G. T., H. Elhefni, T. Salman, I. Waked, M. Abdel-Hamid, N. N. Mikhail, G. Esmat, and A. Fix. 2002. Role of hepatitis C infection in chronic liver disease in Egypt. Am. J. Trop. Med. Hyg. **67**:436–442.

Strimmer, K., and O. G. Pybus. 2001. Exploring the demographic history of DNA sequences using the generalized skyline plot. Mol. Biol. Evol. **18**:2298–2305.

Yesner, D. R. 2001. Human dispersal into interior Alaska: antecedent conditions, mode of colonization, and adaptations. Quat. Sci. Rev. **20**:315–327.

Arndt von Haeseler, Associate Editor

Accepted January 25, 2005