BMC Evolutionary Biology

BioMed Central

Methodology article

Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo

Rainer Opgen-Rhein, Ludwig Fahrmeir and Korbinian Strimmer*

Address: Department of Statistics, University of Munich, Ludwigstr. 33, D-80539 Munich, Germany

Email: Rainer Opgen-Rhein - opgen@stat.uni-muenchen.de; Ludwig Fahrmeir - fahrmeir@stat.uni-muenchen.de; Korbinian Strimmer* - korbinian.strimmer@lmu.de

* Corresponding author

Published: 21 January 2005

BMC Evolutionary Biology 2005, **5**:6 doi:10.1186/1471-2148-5-6

Received: 29 June 2004
Accepted: 21 January 2005

This article is available from: http://www.biomedcentral.com/1471-2148/5/6

Abstract

Background: Coalescent theory is a general framework to model genetic variation in a population. Specifically, it allows inference about population parameters from sampled DNA sequences. However, most currently employed variants of coalescent theory only consider very simple demographic scenarios of population size changes, such as exponential growth.

Results: Here we develop a coalescent approach that allows Bayesian non-parametric estimation of the demographic history using genealogies reconstructed from sampled DNA sequences. In this framework inference and model selection is done using reversible jump Markov chain Monte Carlo (MCMC). This method is computationally efficient and overcomes the limitations of related non-parametric approaches such as the skyline plot. We validate the approach using simulated data. Subsequently, we reanalyze HIV-1 sequence data from Central Africa and Hepatitis C virus (HCV) data from Egypt.

Conclusions: The new method provides a Bayesian procedure for non-parametric estimation of the demographic history. By construction it additionally provides confidence limits and may be used jointly with other MCMC-based coalescent approaches.

Background

The coalescent is a very versatile stochastic model of the genetic variation in a set of sequences sampled from a population. It allows to accommodate a wide range of assumptions about rates and modes of evolution, and of population history [1-5].

As the observed sequence data are positively correlated due to common ancestry, coalescent theory also provides a framework for understanding the relationship between a population's history and its genealogy. For instance, it has long been noted that genealogies of samples taken from exponentially growing populations tend to be "star-like" with short branch lengths near the root of the tree. In contrast, the inter-node distances in genealogies from constant-size populations typically are much more evenly spaced.

Thus, coalescent theory quantifies the imprint that demographic development of a population leaves in the data. While the original theory was outlined for constant population size [1,2], Slatkin and Hudson [6] soon developed a coalescent model for the case of an exponentially growing population. Subsequently, a general approach allowing arbitrary population size variation through time was presented by Griffith and Tavaré [7].

Therefore at least in principle the coalescent model provides a basis for *statistically inferring the demographic history* as a function of time from the sampled sequences [3,8-12] or, alternatively, from the corresponding inferred genealogies [13-15]. In practice, however, application of coalescent theory to this problem has been restricted to very simple demographic scenarios such as constant size, exponential or logistic growth.

Only recently methods have emerged that attempt the completely non-parametric estimation of the demographic function from the data. Polanski et al. proposed an approach based on pairwise distances [16], hence generalizing the method by Slatkin and Hudson [6]. Pybus et al. [14] presented the "skyline plot" method that uses a step-function to approximate the population history obtained from an estimated genealogy. This method was subsequently refined to the "generalized skyline plot" [17] which is essentially a regularized version of the classic skyline plot. If the population size is truly constant through time the generalized skyline plot estimate of population size collapses to the phylogenetic coalescent estimator proposed by Felsenstein [13].

The advantage of the skyline plot over the method suggested by Polanski et al. [16] is that it takes into account the genealogical relationship among the sequences. This helps to decrease bias and improves the efficiency of the resulting estimator compared to methods based on summary statistics and pairwise distances [13]. Unfortunately, the skyline plot approach also has several deficiencies. First, it is unclear how to extend the approach to allow multiple genealogies as input. This is important in order to accommodate phylogenetic error, and to allow non-parametric inference of population history in coalescent approaches that take all possible genealogies into account [7-10]. Second, and perhaps more critical, the (generalized) skyline plot only provides a population size trend rather than a realistic estimate of population size changes, as by construction the population function is modeled by a step function. Moreover, the change-points of this function are fixed at the inter-nodes of the underlying tree.

In this paper we propose a novel framework to non-parametric estimation of the demographic history. This approach relies on Bayesian reversible-jump MCMC inference [18] to obtain a smooth population size function from a given set of genealogies. The new method not only renders many deficiencies of the classic and generalized skyline plot obsolete but it is also computationally efficient, with running times of the algorithm for typical data in the order of minutes on standard PC hardware. The framework has been implemented in the computer language R [19] and incorporated in the R package APE [20].

The remainder of the paper is organized as follows. In the next section we describe the mathematical and statistical theory of the new framework. Subsequently, we apply the method to simulated and biological sequence data and discuss the results. In the last section we briefly outline possible further extensions and related directions of research.

Results

Background in coalescent theory

Basic model

In a pan-mictic population with constant effective population size N_e, where every individual has a single parent, the waiting time w_n until any two of n sampled lineages coalesce is exponentially distributed with rate $r_n = \binom{n}{2} / N_e$ [1,2]. For n sequences there are therefore n - 1 intervals I_n, I_{n-1},..., I_2 with rates r_n, r_{n-1},..., r_2 and interval lengths w_n, w_{n-1},..., w_2. With $T = \sum_{i=2}^{n} w_i$ we denote the time until all samples have reached the most recent common ancestor.

The coalescent model implies that the waiting time to the next coalescent event follows an inhomogeneous Poisson-process with a hazard rate r_n that varies in time t because of the change in the number of lineages. Thus, it is straightforward to also include variable population size in the coalescent simply by using the hazard rate $r_n = \binom{n}{2} / N_e(t)$. From standard theory in survival analysis [21] it follows that the corresponding density for the waiting times is given by

$$\mathrm{P}(\tau_{i+1} - \tau_i \mid \tau_i = r(\tau_{i+1}) \cdot \exp\left[-\int_{v=\tau_i}^{\tau_{i+1}} r(v)dv \right], \quad (1)$$

where τ_i is the time at the beginning of the interval I_i. This is exactly the distribution from the variable population size coalescent

$$\mathrm{P}(w_i \mid \tau_i) = \frac{\binom{i}{2}}{N_e(w_i + \tau_i)} \exp\left[-\int_{v=\tau_i}^{w_i+\tau_i} \frac{\binom{i}{2}}{N_e(v)} dv \right] \quad (2)$$

as developed in [7]. The coalescent model can be further expanded to diploid populations [22] or to include other

effects like selection, recombination or geographical structures [4]. In this paper, however, we focus solely on the coalescent/survival model given by Eq. 2.

Estimation of population size

If the waiting times w_i are known Eq. 2 can be used directly to estimate $N_e(t)$. This is typically done by maximizing the likelihood $L = \prod_{i=2}^{n} P(w_i \mid t_i)$ assuming a simple parametric model for the population size change. For constant population size this has been done in [13], for more complicated scenarios such as logistic growth see, e.g., [14].

In a typical setting, however, the waiting times are themselves estimated from sequence data. In this case the total likelihood function will be a weighted sum of the likelihoods for all possible waiting times, so that in effect the w_i are marginalized out in favor of the actually observed data. In practice exact calculation of this sum is prohibitive, hence one relies on approximating MCMC methods [8-10].

As a shortcut to avoid these computationally very expensive procedures one may also substitute the "true" waiting times by those obtained from inter-node distances of a single estimated gene tree (see, e.g., [23] for an overview of relevant likelihood-based tree inference methods) and proceed as above. Note that the resulting plug-in approximation ignores the uncertainty from estimating the w_i in the inference of demographic parameters. However, this is justifiable if the phylogenetic error is much smaller than the error introduced by the coalescent. This will be the case if sequences are sufficiently long and the substitution rate is comparatively high (a typical example would be virus data).

For non-parametric estimation of population size, Pybus et al. suggested the "skyline plot" [14]. This method assumes a piece-wise constant function for the population size $N_e(t)$ and allows population size changes only at the beginning and end of an interval I_i. The estimated effective population size $\hat{H}_i$ in interval I_i according to the skyline plot is given by the simple relation

$$\hat{H}_n = \hat{w}_n \frac{n(n-1)}{2}. \qquad (3)$$

This is the maximum likelihood estimate under the assumed model of fixed change-points. The "generalized skyline plot" subsequently introduced by Strimmer and Pybus [17] reduces the over-fitting present in the classic skyline plot by applying a simple form of regularization: adjacent intervals that alone are likely to have high stochastic noise are pooled together (cf. Fig. 2b and 2d). Choice of an optimal grouping of intervals (i.e. model selection) is performed by employing a second-order variant of the Akaike criterion [24].

A Bayesian non-parametric approach to estimating demographic history

Outline

In this paper we present a non-parametric approach to infer population size changes in time that overcomes the limitations of previous approaches. More specifically, we develop a non-parametric Bayesian estimator for the function $N_e(t)$ conditioned on observed or sampled inter-node distances $w_n, w_{n-1},..., w_2$ by determining the posterior distribution $P(N_e(t)|w_n, w_{n-1},..., w_2)$. In order to sample the non-parametric demographic function from this posterior we use the reversible jump Markov chain Monte Carlo (rjMCMC) algorithm [18]. As a result, we obtain for any given time t both a point estimate $\hat{N}_e(t)$ – here we choose the posterior median – as well as the associated credible interval (e.g., the lower and upper 2.5% quantiles). If the considered inter-node distances $w_n, w_{n-1},..., w_2$ are fixed and obtained from a single estimated tree, the resulting method is already directly applicable to phylogenetically informative data such as viral sequences (this is the focus of this paper). However, sampling of non-parametric demographic functions can also be combined in a conceptually straightforward fashion with sampling of trees, as outlined below.

Bayesian inference using reversible jump MCMC

In a nutshell, Bayesian inference of a parameter x consists of updating its prior distribution $P(x)$ to a posterior distribution $P(x|D)$ that takes account of the information in the observed data D. The relative evidence of the data for different values of x is summarized in the likelihood $L = P(D|x)$ that accordingly plays a central role in the computation of the posterior via Bayes' theorem

$$P(x|D) = \frac{P(x)P(D \mid x)}{\sum_v P(v)P(D|v)}. \qquad (4)$$

For most realistic problems the posterior distribution cannot be computed analytically, in particular if x is a high-dimensional vector. Instead, one utilizes computational procedures to efficiently draw random samples from the posterior. This in turn allows computation of summary statistics such as the median or the upper and lower 2.5% quantiles. Markov chain Monte Carlo (MCMC) is one particularly useful sampling algorithm as it doesn't require calculation of the sum (or integral) in the nominator of Eq. 4. Briefly, sampling via MCMC is done by constructing a Markov chain with the possible combinations of parameter values as "states", and the desired posterior as its stationary distribution. These properties can be guaran-

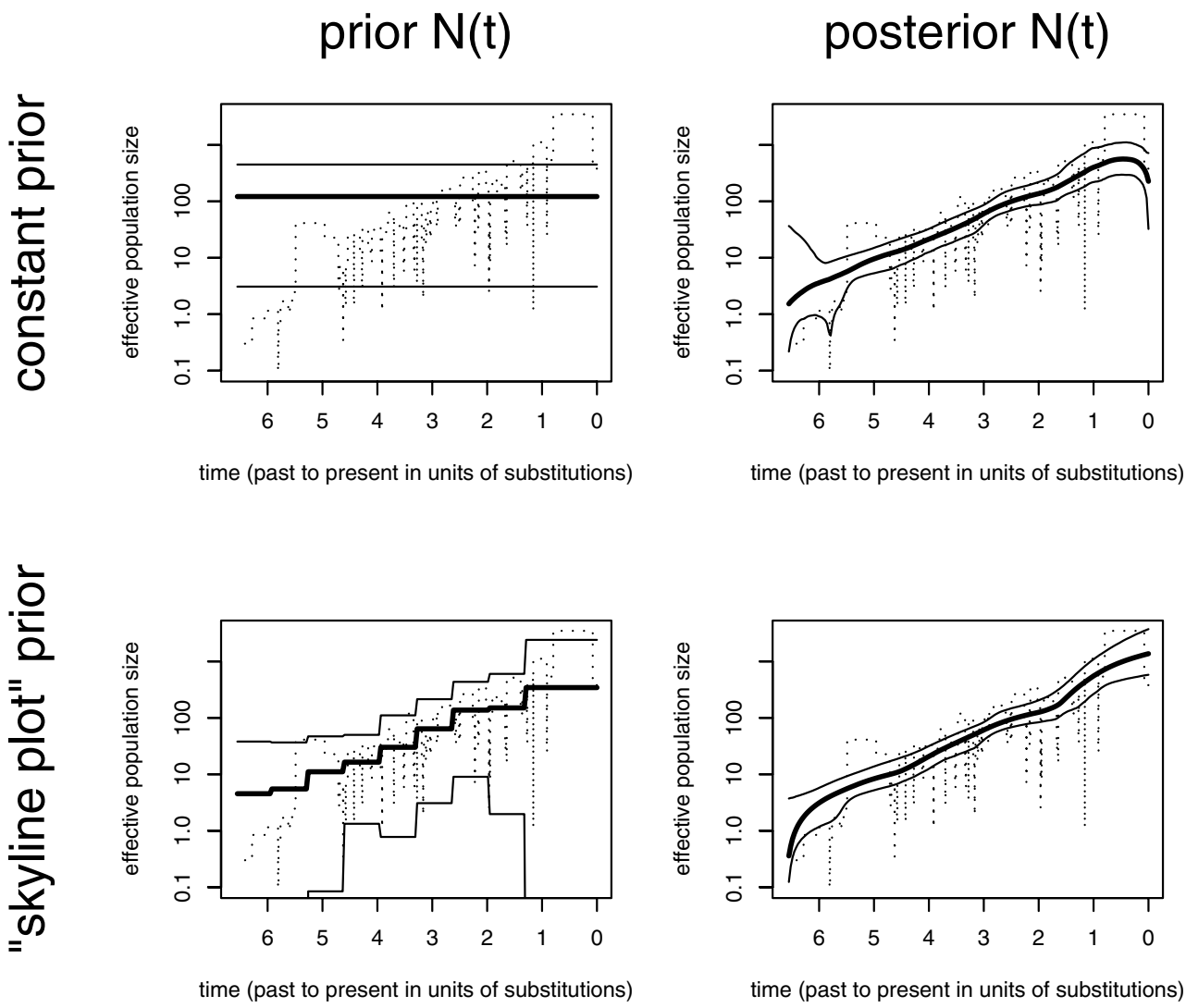

Figure 2
Comparison of prior and posterior demographic functions *Top row*: Bayesian inference using a prior demographic function with constant mean and constant variance (a 95% confidence band is indicated by showing the 2.5% and 97.5% quantiles). *Bottom row*: Bayesian inference using the "skyline plot" prior function.

teed by following certain rules for accepting or rejecting proposed new parameter values. Here we use the Metropolis-Hastings-Green method, i.e. the reversible jump MCMC algorithm [18], that has the advantage of not only allowing changes in the parameters values but also in the dimension of the parameter vector itself. Specifically, if x is the initial state, and $\tilde{x}$ a proposed new state with proposal density $q(\tilde{x})$, then the acceptance probability according to Green [18] is

$$\alpha(x, \tilde{x}) = \min\{1, \mathcal{L} \cdot \mathcal{A} \cdot \mathcal{P} \cdot \mathcal{J}\}, \qquad (5)$$

where $\mathcal{L}$ is the likelihood ratio $P(D|\tilde{x})/P(D|x)$, $\mathcal{A}$ is the prior ratio $P(\tilde{x})/P(x)$, $\mathcal{P}$ is the proposal ratio $q(\tilde{x})/q(x)$, and $\mathcal{J}$ is the determinant of the Jacobian resulting from the potential change of dimension of the parameter vector.

Accordingly, for the application of MCMC to infer the functional form of demographic history a variety of components need to be specified:

• a suitable *parameterization* of the estimated function $N_e(t)$

• the *likelihood function*,

• a *prior distribution* for each considered variable, and,

• rules to construct the *Markov chain* (i.e. acceptance probabilities).

In the following sections we now describe each of these elements in detail. For further general information on the statistical and mathematical background of the MCMC algorithm we refer to the many excellent monographs on this topic (e.g., [25]).

Parameterization of $N_e(t)$

In our suggested procedure we approximate the sampled demographic history $N_e(t)$ by a piecewise linear function. This spline of first order degree consists of a first node at position $a_0 = 0$ and height h_0, followed by k internal supporting nodes at $(a_1; h_1)$, $(a_2; h_2)$,..., $(a_k; h_k)$, and a terminal node at $a_{k+1} = T = \sum_{i=2}^{n} w_i$ with height h_{k+1} Hence, the spline is defined for all $t \in [0, T]$, and for any given k the it contains k free position parameters and $k + 2$ free height parameters. Note that, unlike in the skyline plot, we do not constrain the change-points $a_1,..., a_k$ to lie on the grid points defined by the inter-node distances w_i. Moreover, we also allow that the number of internal nodes k changes during sampling of the population function from the posterior. Hence, k is technically a hyper-parameter that controls the roughness of the resulting spline. As will be clear from the outline of the MCMC algorithm below, note that the final point estimate $\hat{N}_e(t)$ obtained from posterior sampling will be a mixture of linear splines (i.e. a smooth and possibly nonlinear function) rather than a single spline.

Likelihood function

The likelihood L employed in our procedure is the product of the densities of the waiting times between subsequent coalescence events, i.e. $L = \prod_{i=2}^{n} \mathrm{P}(w_i \mid \tau_i)$. This function depends via Eq. 2 on the effective population size $N_e(t)$, and hence indirectly on the spline parameters a_i, h_i and k. Because $N_e(t)$ is represented by a linear spline, calculation of the likelihood can be done in a computationally efficient fashion.

Prior distributions

Number of change-points

Following [18] we employ a truncated Poisson-distribution as the prior distribution for k, i.e.

$$\mathrm{P}(k) = \begin{cases} \frac{1}{c}\frac{\lambda^k}{k!}e^{-\lambda} & \text{for } k \le k_{\max} \\ 0 & \text{otherwise,} \end{cases} \qquad (6)$$

where c is a normalizing constant to ensure that $\mathrm{P}(k)$ is a proper distribution. For the hard upper limit of the number of change-points we use $k_{\max} = 30$. The parameter λ acts as a smoothing parameter, set in a typical analysis to about $\lambda = 0.1$ - 1.0.

As an alternative to using a fixed λ we also suggest a hierarchical Bayes approach where λ is drawn from a Gamma distribution

$$\mathrm{Gamma}(\lambda|a,b) = \frac{1}{b^a \Gamma(a)} e^{-\lambda/b} \lambda^{a-1}, \qquad (7)$$

with some shape parameter a and scale parameter b (for instance, $a \approx 0.5$ and $b \approx 2$ so that $\mathrm{E}(\lambda) = ab \approx 1$ and $\mathrm{Var}(\lambda) = ab^2 \approx 2$).

Positions

We assume that the internal nodes of the spline are *a priori* uniformly distributed in the interval $[0, T]$. As a simple trick to avoid very small inter-node distance we generate $2k + 1$ random variables, and set the change-points $a_j = z_{[2j]}$ for $j = 1,..., k$. The corresponding joint density is

$$\mathrm{P}(a_1,...,a_k) = (2k+1)!\left(\frac{1}{T}\right)^{2k+1} \prod_{j=0}^{k}\left(a_{j+1} - a_j\right) \qquad (8)$$

with $a_0 = 0$ and $a_{k+1} = T$.

Heights

As prior distributions for the heights h_i we assume a Gamma distribution

$$\mathrm{Gamma}(h_i|\alpha_i, \beta_i) \qquad (9)$$

which ensures that sampled heights are always positive. The parameters α_i and β_i determine the *a priori* mean and variance of height h_i. More generally, one can also allow fully time-dependent prior parameters $\alpha(t)$ and $\beta(t)$. This is particularly advisable if the population size is known in advance to be subject to large changes in time.

In a strict Bayesian approach, the choice of the prior distribution for the heights is completely external to the observed data. One simple possibility would, e.g., be to assume an arbitrary constant for the mean and variance. However, we recommend to follow a more pragmatic "empirical Bayes" route and to use the data at hand (or some other related data set) to obtain an informed guess about the prior heights. For example, an assumed constant population size as prior mean could be estimated using the method by Felsenstein [13]. Another possibility is to employ the skyline plot as a prior mean estimate (this is the default in our program).

However, note that in practice the actual choice of prior height distribution seems to matter only little for estimating the posterior demographic function (see Figure 2 and the section on simulated data below). Only when there are few coalescent events per unit of time will the posterior estimate of the demographic function be dominated by the prior.

Construction of the Markov chain

There are four different possibilities to change the state defined by the parameters c_i, h_i, and k of the spline describing the effective population size $N_e(t)$:

1. varying the position of a change-point (i.e. internal node),

2. changing the height at a certain change-point,

3. generating a new change-point ("birth" step), and

4. deleting an existent change-point ("death" step).

Let η_k, π_k, b_k, and d_k the probabilities of the four moves given k, with $\eta_k + \pi_k + b_k + d_k = 1$. In order to satisfy the requirement of detailed balance in the corresponding Markov chain the probabilities of birth and death steps (b_k and d_k) need to be synchronized [18]. This can be achieved, e.g., by setting

$$b_k = c\min\left\{1, \frac{\mathrm{P}(k+1)}{\mathrm{P}(k)}\right\} \qquad (10)$$

and

$$d_{k+1} = c\min\left\{1, \frac{\mathrm{P}(k)}{\mathrm{P}(k+1)}\right\}, \qquad (11)$$

where c is chosen so that $b_k + d_k < 0.9$ for all k.

Next, we describe the individual procedures to propose and accept one of the above four moves as implemented in our program.

Height change

First, a height h_j is selected out of the $k + 2$ existing heights with probability $\frac{1}{k+2}$. Second, a new height $\tilde{h}_j$ is generated by $\tilde{h}_j = h_j \exp(z)$, where z is a uniformly distributed random variable on $\left[-\frac{1}{2}, \frac{1}{2}\right]$. Third, the new height is accepted with probability

$$\alpha_H(x, \tilde{x}) = \min\{1, \mathcal{L}\cdot\left(\frac{\tilde{h}_j}{h_j}\right)^{\alpha} \exp\left(-\beta\left(\tilde{h}_j - h_j\right)\right)\}, \qquad (12)$$

where α and β are from the prior distribution and $\mathcal{L}$ denotes the ratio of the likelihood of the new state $\tilde{x}$ (with modified height) and the likelihood of the current state x.

Position move

First, a change-point a_j is chosen randomly with probability $\frac{1}{k}$. Second, its new position $\tilde{a}_j$ within $[a_{j-1}, a_{j+1}]$ is determined by drawing from the corresponding uniform distribution. Third, $\tilde{a}_j$ is accepted with probability

$$\alpha_p(x, \tilde{x}) = \min\{1, \mathcal{L}\cdot\frac{\left(a_{j+1} - \tilde{a}_j\right)\left(\tilde{a}_j - a_{j-1}\right)}{\left(a_{j+1} - a_j\right)\left(a_j - a_{j-1}\right)}\}. \qquad (13)$$

Birth step

First, the position a^* of the new change-point is found by uniformly drawing from $(0, L)$, and let the neighboring nodes left and right of a^* have positions a_j and a_{j+1}. Second, the corresponding new height h^* is generated by randomly disturbing the current height $N_e(a^*)$ on the position a^* according to $N_e(a^*) + zN_e(a^*)$ where z is a uniformly distributed random variable on the interval $\left[-\frac{1}{2}, \frac{1}{2}\right]$. Note that the birth step increases the dimension of the parameter vector from $2k + 2$ to $2k + 4$ as a new change-point and a new height are generated.

The corresponding acceptance probability of the birth step is computed according to Eq. 5 with likelihood and prior ratios as above, and with proposal ratio

$$\mathcal{P} = \frac{d_{k+1}}{b_k}\frac{T}{(k+1)} \qquad (14)$$

and Jacobi determinant

$$\mathcal{J} = \frac{\left(a^* - a_j\right)\left(h_{j+1} - h_j\right)}{a_{j+1} - a_j} + h_j. \qquad (15)$$

Death step

This is the inversion of the birth step and consists of removing a change-point. First, a^* chosen from $a_1,\ldots, a_k$ with probability $\frac{1}{k}$. Second, the corresponding height h^* is also removed from the vector of spline parameters. The acceptance probability for the death step is

$$\alpha_D(x, \tilde{x}) = \min\{1, (\mathcal{L}\cdot\mathcal{A}\cdot\mathcal{P}\cdot\mathcal{J})^{-1}\}, \qquad (16)$$

where the proposal ratio and the Jacobi determinant is the same as for the birth step.

Computation of estimated $N_e(t)$ and associated confidence intervals

In order to obtain an estimate of the effective population size in time we now proceed as follows. First, the Markov chain is started with an initial state that corresponds to a completely flat demographic function, i.e. $N_e(t) = c$, where c is some rough estimate of population size, and $k = 0$. Second, 100,000 repeats of the MCMC algorithm are performed, of which the first 5,000 are ignored to allow for a "burn-in" period.

Third, the remaining samples are thinned out by a factor of 1:50 to remove auto-correlation. As a result, 1900 independent samples from the joint posterior of the spline parameters a_i, h_i and k are obtained.

Subsequently, in order to obtain a point estimate $\hat{N}_e(t)$ and associated confidence bands we compute the distribution of the effective population size at a number of fixed equidistant time points $t_1, t_2,\ldots, t_{1000} \in [0, T]$. Finally, we report as summary statistics the corresponding median and the lower and upper 2.5% quantiles.

Extension to multiple genealogies

In this paper we have introduced non-parametric sampling of demographic histories assuming a fixed underlying genealogical tree (or equivalently, a fixed set of internode distances $w_n, w_{n-1},\ldots, w_2$.)

However, in our approach – unlike previous non-parametric methods such as the skyline plot – it is also conceptually straightforward to incorporate phylogenetic error. This can be done by joint sampling of trees and demographies according to the following simple algorithm:

1. Given sequence data D, sample a tree G^* with clock-like branch lengths (see, e.g, refs. [8,9,11,12,26] for suitable methods).

2. Use the method described in this paper to sample the demographic function conditioned on the inter-node distances w_i^* from G^*.

3. Repeat steps 1 and 2 to obtain the posterior distribution for the population size function, now conditioned on D rather than on some given $w_n, w_{n-1},\ldots, w_2$.

Note that each sampled tree may have a different depth $T^* = \sum_{i=2}^{n} w_i^*$. This means that the interval [0, T] for the prior (and posterior) height distribution has to be set in advance (and independent of the T^*). For the case of $0 < t < T^*$ sampling of heights then proceeds as described above, while for $T^* < t < T$ – the region with no data from a given sampled tree – the heights are simply drawn from the respective prior distribution.

Discussion

In order to test the potential of the proposed reversible jump MCMC algorithm we first applied it to synthetic data simulated according to various demographic scenarios. Subsequently, we reanalyzed two viral data sets from Central Africa and Egypt.

Computer program

The proposed framework has been implemented by us for the case of a single underlying genealogy. The program is written in the statistical computer language R [19] and is incorporated in recent versions of the R package APE [20].

To install the APE package, simply run the R program, and enter at the R prompt

```
install.packages("ape")
```

This downloads the APE package from the Internet. The proposed reversible jump MCMC approach is implemented in the function "mcmc.popsize" of which an extensive description along with examples can be obtained online by typing

```
library("ape")

help(mcmc.popsize)
```

into the R command window. The APE package also includes routines for plotting the inferred population

function (e.g., all figures in this paper were prepared with APE).

Note that the use of this R program is only valid if the phylogenetic error is low – this is typically the case when the evolutionary rate is high and the available sequences are long (e.g. viral data). If the phylogenetic error is not negligible compared to the coalescent error, please use software such as BEAST [27].

Simulated data

In the simulation setup we followed Pybus et al. [14] and Strimmer and Pybus [17]. Specifically, we performed simulations assuming constant population size ($N_e(t)$ = 100) as well as exponential population growth ($N_e(t)$ = l000e^{-t}), using 25 and 100 sampled lineages, respectively. To estimate the population size function we employed the proposed MCMC algorithm and the classic and generalized skyline plot. In the former the smoothing parameter λ was drawn from the hierarchical model with default parameters (a = 0.5 and b = 2).

Figure 1 shows the results from a typical run of the simulations. The top row illustrates the case of constant population size, whereas the bottom row demonstrates exponential growth. On the left in Figure 1, top row, the true underlying constant population size is shown (the thick dashed line), together with the estimate provided by the classic skyline plot. On the right, this is contrasted with the estimate obtained by using our reversible jump MCMC algorithm. Clearly, the median of the posterior distribution of $N_e(t)$ is a very good point estimator of the true demographic history. In addition, the 95% confidence band is also automatically obtained by the MCMC method. Interestingly, it can be immediately seen that the uncertainty in $N_e(t)$ increases with a growing distance from the present. This simply reflects the fact that near the root of the tree for constant population size there are only few coalescent events.

In Figure 1, bottom row, an example for a simulation with an exponentially growing population is shown. As for the constant population, the rjMCMC algorithm is capable of recovering the original population size function (shown as thick dashed line) complete with confidence bands, whereas the skyline plot contains a large amount of stochastic noise, and only provides a rough exploratory picture of the population size changes.

In Figure 2 the influence of the choice of prior demographic function on the final posterior estimate is investigated using further simulations of an exponentially growing population. The left column depicts the prior distributions (specifically the 2.5%, 50% and 97.5% quantiles for each time point) for two typical cases: a constant prior function (= constant population size with constant variance), and the "skyline plot" prior function (= time dependent piecewise- constant population size and variance). The right column of figure 2 presents the corresponding posterior distributions as obtained with the present rjMCMC approach. The results for both cases are very similar. This indicates that there is sufficient signal in the data to make the posterior demographic function (almost) independent from the choice of prior distribution. Note that only near the left and right end of the investigated time intervals there are some slight differences. These can be explained by the lack of data points near the borders.

HIV-1 in Central Africa

Next, we applied our method to infer the demographic history from a set of HIV-1 sequences from Central Africa. These data was originally used by Vidal et al. [28] who examined the genetic diversity of HIV-1 type M in this region. Further detailed analysis can be found in Rambaut et al. [29] and Yusim et al. [30]. Here we use the reconstructed phylogeny of Yusim et al. with which Strimmer and Pybus also estimated the demographic history by means of the generalized skyline plot [17].

Figure 3 shows the result of the analysis with the reversible jump MCMC algorithm compared with the predictions from the classic and generalized skyline plots. As in Yusim et al. [30] an evolutionary rate of 0.0023 substitutions per year was assumed to convert the time axis into units of years. The first row of Figure 3 displays the tree of Yusim et al. [30] and the corresponding classic skyline plot. The latter exhibits a large amount of noise, nevertheless the main demographic signal is clearly visible in the graph. In contrast, in Figure 3c (second row) the effective population size as estimated by the rjMCMC algorithm is displayed. The thick line shows the median and the thin lines the 95% confidence interval. Especially in the middle part of the figure, where most of the data is located, the confidence interval is very narrow, indicating a stable estimation of the demographic history. Also note that for this data the average number of change-points in the MCMC run was k = 9.25, i.e. the estimated effective degree of freedom is much less than that implicitly assumed in the classic skyline plot.

A comparison with the generalized skyline plot [17] is shown in Figure 3d. This demonstrates that the generalized skyline plot, in contrast to its classic cousin, provides a very good noise-reduced approximation to the demographic history as estimated by the reversible jump MCMC approach. However, especially near the present the step function employed in the generalized skyline plot leads to unrealistic jumps in the population size that are

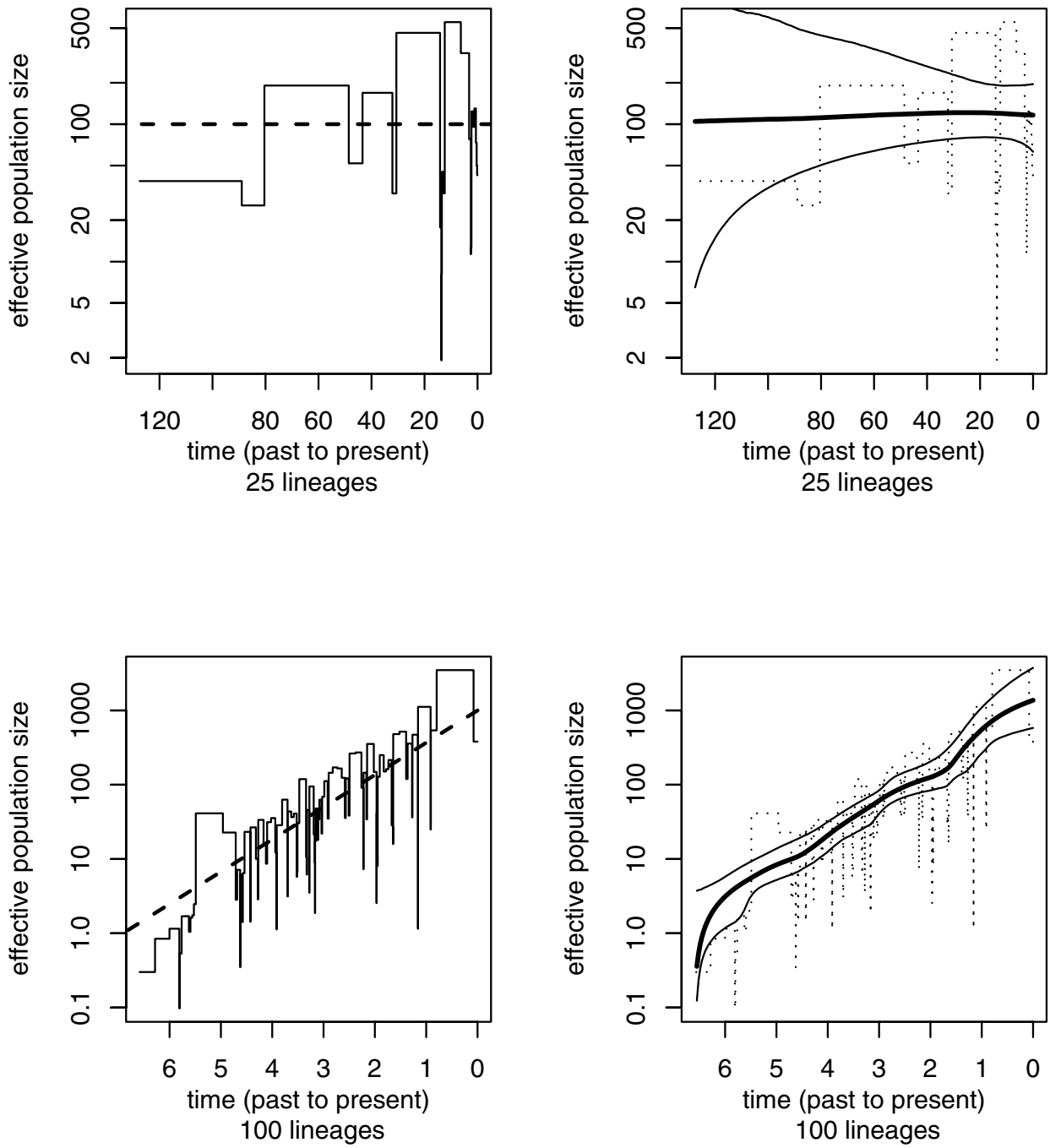

Figure 1

Simulated data *Top row*: Example of a simulation with constant population size: (left) true demographic history (dashed line) and estimate obtained with the classic skyline plot; (right) point estimate obtained with rjMCMC and 95% confidence band. *Bottom row*: Example with exponential population growth: (left) true population growth and classic skyline plot; (right) results from rjMCMC approach.

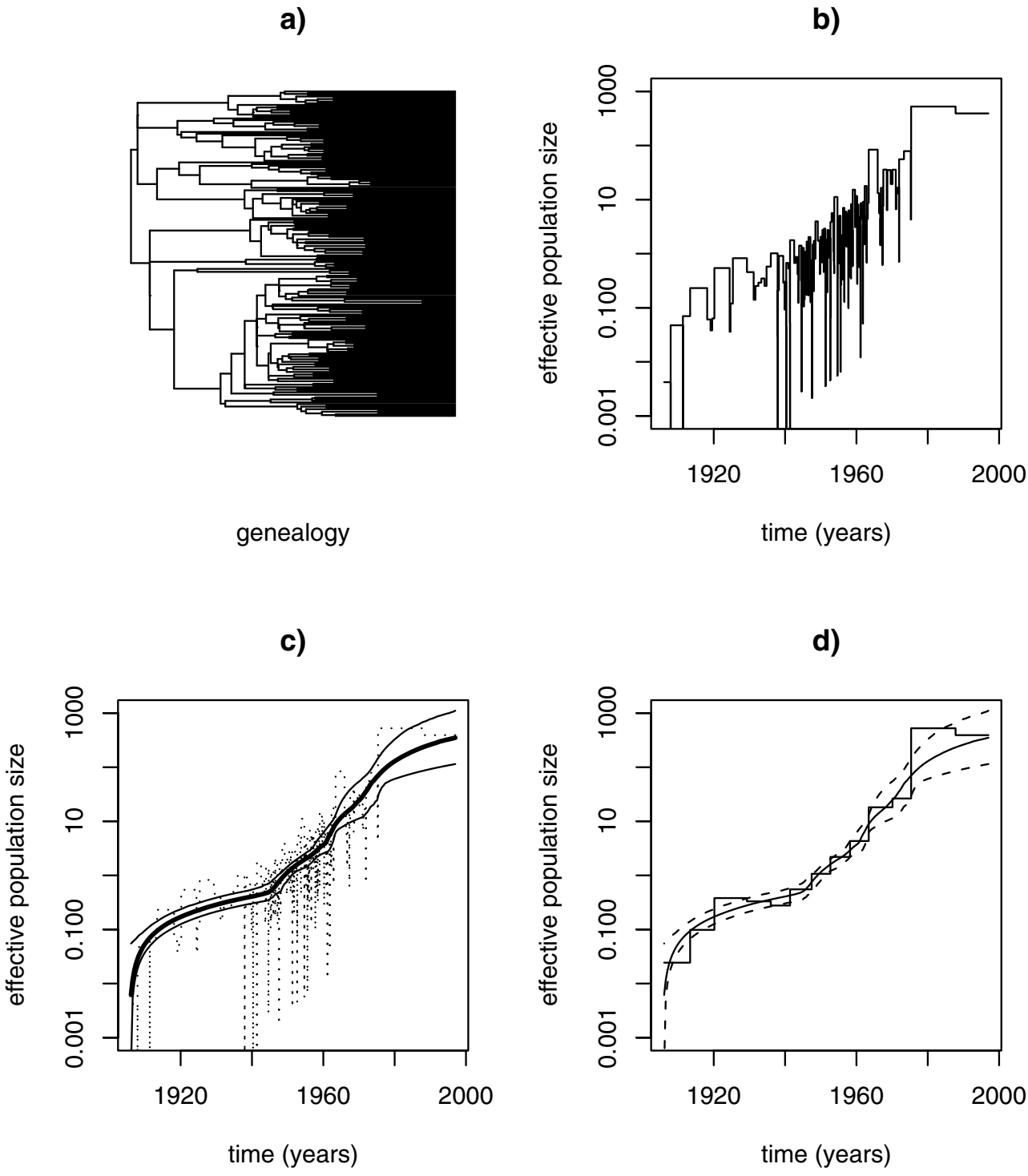

Figure 3
HIV-1 in Central Africa *Top row*: a) underlying genealogy; b) classic skyline plot. *Bottom row*: c) population size function estimated with rjMCMC and corresponding 95% confidence band; d) comparison rjMCMC versus generalized skyline plot.

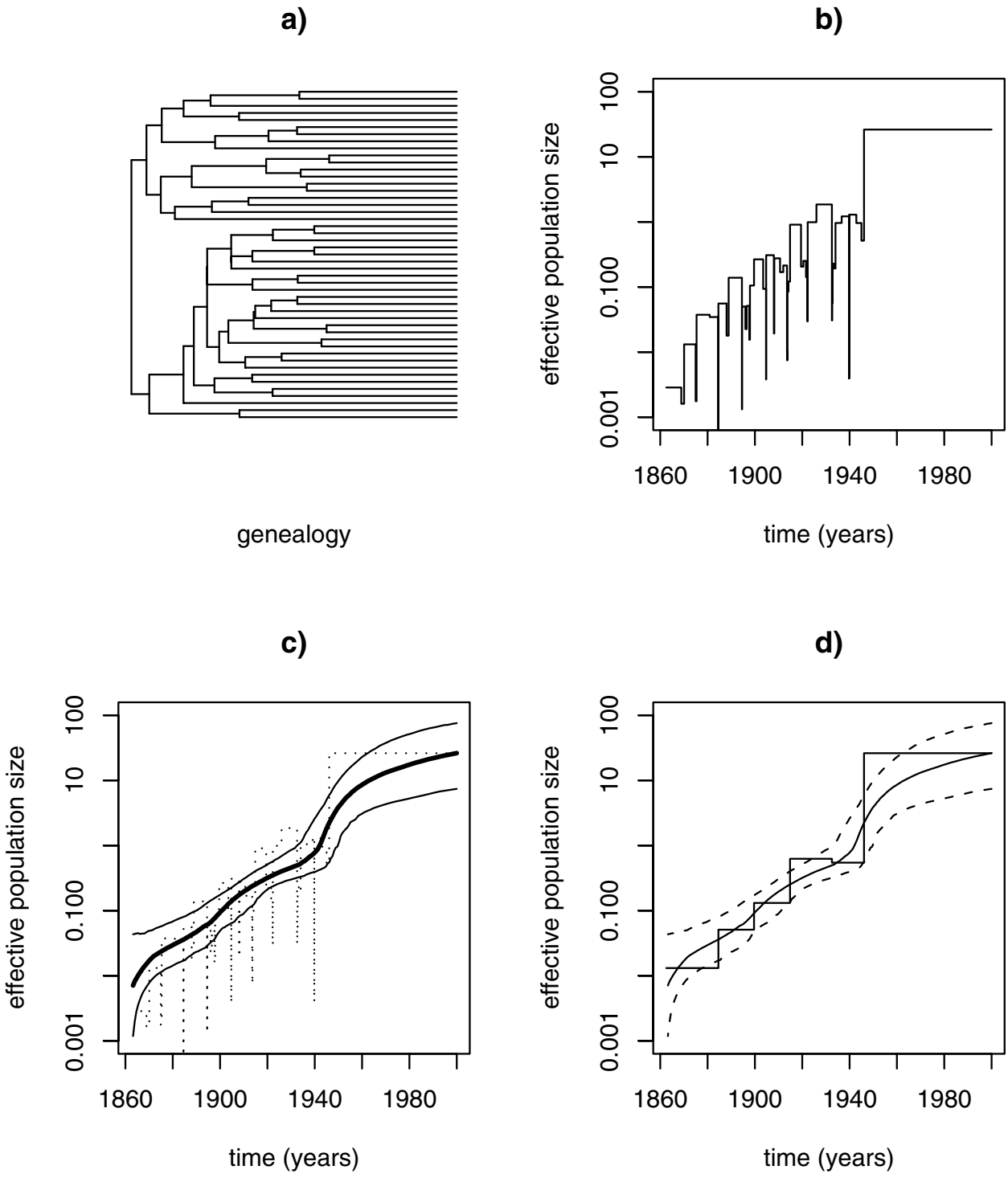

Figure 4
HCV in Egypt *Top row*: a) underlying reconstructed genealogy; b) classic skyline plot. *Bottom row*: c) population size function estimated with rjMCMC and corresponding 95% confidence band; d) comparison rjMCMC versus generalized skyline plot.

not present in the smooth estimate provided by the proposed MCMC method.

HCV in Egypt

In Egypt 10%-20% of the general population are infected with the Hepatitis C virus (HCV) [31]. This endemicity seems mainly to be caused by percutaneous medical procedures such as needle injections that took place during a countrywide health campaign between 1964 and 1982 against schistosomiasis. In order to investigate this phenomenon blood samples were obtained from various regions of Egypt and used to study the epidemic history of Hepatitis C. For instance, Tanaka et al. [32] analyzed the molecular evolution of HCV genotype 4a. Specifically, they utilized 47 sequences (AF217800-AF217812 [31] and AB103424-AB103457 [32]) from the NS5B region of the HCV subtype 4a to reconstruct the respective phylogeny, and subsequently applied the skyline plot method to infer the demographic history.

We repeated their analysis with the reversible jump MCMC approach developed in this paper. We downloaded the sequence data from the HCV sequence database [33] and inferred the corresponding maximum-likelihood genealogy using the TREEFINDER program [34]. This tree is depicted in Figure 4a, next to the demographic history estimated from it by the classic skyline plot (Figure 4b). In the bottom of the figure we show the estimated population size function and its 95% confidence bands as obtained by our rjMCMC method (Figure 4c) and we also compare our results with those of the generalized skyline plot (Figure 4d). For the generating the time axis in these plots we assumed an evolutionary rate of 0.00045 substitutions per year.

Generally, the star-like shape of the inferred tree already is indicative of exponential growth. This is confirmed by both the skyline plot as well as by our analysis (Figure 4d). Moreover, it can be seen that around 1940 the growth rate increased (i.e. the slope of $N_e(t)$ in the log-plot changes). Near the present, the rate decreased again. Also note the broadening of the confidence interval since 1940 which reflects the sparsity of available observations. This implies that the claim of Tanaka et al. [32] that the demographic history recently changed back to constant population size after an exponential growth is not firmly backed by the data. For further biological analysis of the HCV data we refer to Pybus et al. [35].

Conclusions

We have presented a new approach to non-parametric inference of demographic history from an inferred genealogy. This method is based on reversible jump MCMC sampling of the population size function $N_e(t)$. Unlike its predecessors, the classic and generalized skyline plots, it returns a smooth and realistic estimate of the demographic history and thus overcomes the constraints due to assuming a step function. Moreover, it automatically provides confidence limits. Nevertheless, the procedure is still computationally fast and can be run on any standard PC hardware.

In our examples we demonstrated the advantage of non-parametric estimation of demographic history. Parametric estimation always assumes a certain functional form of population growth which may lead to problematic statements (cf. the HCV data set), in particular if the confidence bands of the estimated function $N_e(t)$ are not taken into account.

From the methodological point of view, model selection via rjMCMC has the advantage that the effective dimension, i.e. the degree of smoothing, is automatically chosen in a data-driven manner. There is only one parameter (λ) that controls the *a priori* degree of smoothing, and this is adjusted accordingly by the investigated data. In addition, a further advantage of our MCMC approach is that – in contrast to the skyline plot – at least in principle it is straightforward to incorporate it in more general MCMC sampling schemes that also take account of the uncertainty in the genealogy.

During the referee process we have learned that the authors of the software package BEAST [27] have developed a similar non-parametric method to Bayesian coalescent inference of population history (A.J. Drummond et al., in preparation). We plan to work with Dr. Drummond to make available in BEAST joint sampling of sampling of demographic histories and of trees. This would combine the present rjMCMC approach and the method developed by Drummond and colleagues.

Authors' contributions

This paper summarizes the main results from a master's thesis of R.O. supervised by K.S. and L.F. Accordingly, K.S. and L.F. jointly devised the project and R.O. carried out all analyses and simulations. All authors participated in the development of methodology. R.O. and K.S. wrote the manuscript. All authors approved of the final version.

Acknowledgements

We are grateful for financial support from the Deutsche Forschungsgemeinschaft (DFG) in the Emmy Noether program (R.O. and K.S.) and from the SFB 386 (L.F). We thank G??nter Ra??er and Leonhard Held for valuable comments and discussion and Juliane Sch??fer for critical reading of the manuscript.

References

1. Kingman JFC: **The coalescent.** *Stoch Proc Applns* 1982, **13:**235-248.
2. Kingman JFC: **On the genealogy of large populations.** *J Appl Probab* 1982, **19A:**27-43.

3. Donnelly P, Tavaré S: **Coalescents and genealogical structure under neutrality.** *Annu Rev Genet* 1995, **29:**401-421.
4. Nordborg M: **Coalescent Theory.** In *Handbook of Statistical Genetics* Edited by: Balding D, Bishop M, Cannings C. Chichester: Wiley; 2001:179-212.
5. Hein JJ, Schierup MH, Wiuf CH: *Gene Genealogies, Variation and Evolution* Oxford: Oxford University Press; 2004.
6. Slatkin M, Hudson RR: **Pairwise comparison of mitochondrial DNA sequences in stable and exponentially growing populations.** *Genetics* 1991, **129:**555-562.
7. Griffith RC, Tavaré S: **Sampling theory for neutral alleles in a varying environment.** *Phil Trans R Soc Lond B* 1994, **344:**403-410.
8. Kuhner MK, Yamato J, Felsenstein J: **Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling.** *Genetics* 1995, **140:**1421-1430.
9. Kuhner MK, Yamato J, Felsenstein J: **Maximum likelihood estimation of population growth rates based on the coalescent.** *Genetics* 1998, **149:**429-434.
10. Stephens M, Donnelly P: **Inference in molecular population genetics (with discussion).** *J R Statist Soc B* 2000, **62:**605-655.
11. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W: **Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data.** *Genetics* 2002, **161:**1307-1320.
12. Rannala B, Yang Z: **Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci.** *Genetics* 2003, **164:**1645-1656.
13. Felsenstein J: **Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates.** *Genet Res* 1992, **59:**139-147.
14. Pybus OG, Rambaut A, Harvey PH: **An integrated framework for the inference of viral population history from reconstructed genealogies.** *Genetics* 2000, **155:**1429-1437.
15. Wiuf C: **Inferring population history from genealogical trees.** *J Math Biol* 2003, **46:**241-264.
16. Polanski A, Kimmel M, Chakraborty R: **Application of a time-dependent coalescence process for inferring the history of population size changes from DNA changes.** *Proc Natl Acad Sci USA* 1998, **95:**5456-5461.
17. Strimmer K, Pybus OG: **Exploring the demographic history of a sample of DNA sequences using the generalized skyline plot.** *Mol Biol Evol* 2001, **18:**2298-2305.
18. Green PJ: **Reversible jump Markov chain Monte Carlo computation and Bayesian model determination.** *Biometrika* 1995, **82:**711-732.
19. R Development Core Team: **R: A language and environment for statistical computing.** *R Foundation for Statistical Computing, Vienna, Austria* 2004 [http://www.R-project.org]. [ISBN 3-900051-07-0]
20. Paradis E, Claude J, Strimmer K: **APE: Analyses of phylogenetics and evolution in R language.** *Bioinformatics* 2004, **20:**289-290.
21. Fahrmeir L, Hamerle A, Tutz G, (Eds): *Multivariate statistische Verfahren* 2nd edition. Berlin: Walter de Gryter & Co; 1996.
22. Rosenberg NA, Nordborg M: **Genealogical Trees, Coalescent Theory and the Analysis of Genetic Polymorphisms.** *Nat Rev Genet* 2002, **3:**380-390.
23. Felsenstein J: *Inferring Phylogenies* Sunderland, MA: Sinauer Associates; 2004.
24. Burnham KP, Anderson DR: *Model Selection and Inference: A Practical Information -Theoretic Approach* New York: Springer Verlag; 1998.
25. Gilks W, Richardson S, Spiegelhalter D, (Eds): *Markov Chain Monto Carlo in Practice* 4th edition. London: Chapman and Hall; 1996.
26. Larget B, Simon DL: **Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees.** *Mol Biol Evol* 1999, **16:**750-759.
27. Drummond AJ, Rambaut A: **BEAST: Bayesian Evolutionary Analysis Sampling Trees.** [http://evolve.zoo.ox.ac.uk/beast/].
28. Vidal N, Peeters M, Mulanga-Kabeya C, Nzilambi N, Robertson D, Ilunga W, Sema H, Tishimanga K, Bongo B, Delaporte E: **Unprecedented degree of HIV-1 group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa.** *J Virol* 2000, **74:**10498-10507.
29. Rambaut A, Robertson DL, Pybus OG, Peeters M, Holmes EC: **Phylogeny and the origin of HIV-1.** *Nature* 2001, **410:**1047-1048.
30. Yusim K, Peeters M, Pybus OG, Bhattacharya T, Delaporte E, Mulanga C, Muldoon M, Theiler J, Korber B: **Using HIV-1 sequences to infer historical features of the AIDS epidemic and HIV evolution.** *Phil Trans R Soc Lond B* 2001, **356:**855-866.
31. Ray SC, Arthur RR, Carella A, Bukh J, Thomas DL: **Genetic Epidemiology of Hepatitis C Virus throughout Egypt.** *J Infect Dis* 2000, **182:**698-707.
32. Tanaka Y, Agha S, Saudy N, Kurbanov F, Orito E, Kato T, Abo-Zeid M, Khalaf M, Miyakawa Y, Mizokami M: **Exponential Spread of Hepatitis C Virus Genotype 4a in Egypt.** *J Mol Evol* 2004, **58:**191-195.
33. Kuiken C, Yusim K, Boykin L, Richardson R: **The Los Alamos hepatitis C sequence database.** *Bioinformatics* 2005 in press. http://hcv.lanl.gov
34. Jobb G, von Haeseler A, Strimmer K: **TREEFINDER: A Powerful Graphical Analysis Environment for Molecular Phylogenetics.** *BMC Evolutionary Biology* 2004, **4:**18.
35. Pybus OG, Drummond AJ, Nakano T, Robertson B, Rambaut A: **The epidemiology and latrogenic transmission of Hepatitis C virus in Egypt: a Bayesian coalescent approach.** *Mol Biol Evol* 2003, **20:**381-387.