DOI: 10.1534/genetics.110.125260

Joint Inference of Microsatellite Mutation Models, Population History and Genealogies Using Transdimensional Markov Chain Monte Carlo

Chieh-Hsi Wu* and Alexei J. Drummond*,†,‡,[1]

*Bioinformatics Institute, †Allan Wilson Centre for Molecular Ecology and Evolution and ‡Department of Computer Science, University of Auckland, Auckland 1001, New Zealand

Manuscript received November 16, 2010
Accepted for publication February 18, 2011

ABSTRACT

We provide a framework for Bayesian coalescent inference from microsatellite data that enables inference of population history parameters averaged over microsatellite mutation models. To achieve this we first implemented a rich family of microsatellite mutation models and related components in the software package BEAST. BEAST is a powerful tool that performs Bayesian MCMC analysis on molecular data to make coalescent and evolutionary inferences. Our implementation permits the application of existing nonparametric methods to microsatellite data. The implemented microsatellite models are based on the replication slippage mechanism and focus on three properties of microsatellite mutation: length dependency of mutation rate, mutational bias toward expansion or contraction, and number of repeat units changed in a single mutation event. We develop a new model that facilitates microsatellite model averaging and Bayesian model selection by transdimensional MCMC. With Bayesian model averaging, the posterior distributions of population history parameters are integrated across a set of microsatellite models and thus account for model uncertainty. Simulated data are used to evaluate our method in terms of accuracy and precision of θ estimation and also identification of the true mutation model. Finally we apply our method to a red colobus monkey data set as an example.

MICROSATELLITES, also called short tandem repeats (STRs), are repetitions of a DNA sequence motif with length between 1 and 6 bp. Because they are abundant, widely distributed in the genome, and highly polymorphic, microsatellites have become one of the most popular genetic markers for making inferences on molecular evolution and population genetics (Shikano *et al.* 2010; Spong *et al.* 2010).

Unequal crossing over (Smith 1976; Richard and Pâques 2000) and replication slippage (Levinson and Gutman 1987) are the two main mechanisms proposed that potentially provide an explanation for the high mutation rate of microsatellites. The study by Levinson and Gutman (1987) using *Escherichia coli* showed that replication slippage is the predominant mutation mechanism of microsatellite DNA. Replication slippage occurs when the replicating strand and the template strand disassociate and then realign out of register, forming a loop in one of the strands. If the process of replication continues, a loop formed by the replicating strand gives rise to an insertion while that by the template strand results in a deletion.

The simplest microsatellite model is the stepwise mutation model (SMM) proposed by Ohta and Kimura (1973), which states that the length of the microsatellite increases or decreases by 1 repeat unit at a rate independent of the microsatellite length. Although the SMM has been employed to devise commonly used statistics for estimating genetic divergence (Slatkin 1995) and effective population size (Wehrhahn 1975), the model has some drawbacks. Under the SMM, there is no stationary distribution and under this process the repeat length will eventually grow arbitrarily long, which is inconsistent with empirical microsatellite length distributions from genomic data (Kruglyak *et al.* 1998). Moreover the SMM ignores various properties of microsatellite mutation that have been observed in empirical data. Many different models have been developed in attempts to capture some of these properties.

Observations from many studies support the fact that longer microsatellites have a higher mutation rate (Goldstein and Clark 1995; Wierdl *et al.* 1997; Schlötterer *et al.* 1998). A longer microsatellite allele has more locations for potential slippage errors and hence possesses a greater chance of experiencing a mutation event during replication, as demonstrated by Streisinger and Owen (1985) using bacteriophage T4. This is the motivation behind rate-dependent models such as the proportional slippage model (Kruglyak *et al.* 1998) and others (Calabrese *et al.* 2001; Sibly

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.110.125260/DC1.

Available freely online through the author-supported open access option.

[1] *Corresponding author:* Department of Computer Science, University of Auckland, Auckland 1001, New Zealand. E-mail: alexei@cs.auckland.ac.nz

et al. 2001), which describe the mutation rate as a polynomial function of length in repeat units.

Another property is mutational bias, which exists when the probability of expansion and contraction is unequal for a mutation event. Evidence for this phenomenon has been found in genomes of several species including humans (Rubinsztein *et al.* 1999), which exhibit a preference for expansion, and the bacterium *Mycoplasma gallisepticum* (Metzgar *et al.* 2002), which tends to contract. Models proposed by Calabrese and Durrett (2003) and Walsh (1987) have accounted for this rate asymmetry (see original citations for the stationary distributions of these models).

In one-phase models, a mutation leads to expansion or contraction of the microsatellite by 1 repeat unit only. However, empirical data suggest that mutations can occasionally result in a change in the microsatellite length of >1 unit. According to the two-phase model (TPM), proposed by Di Rienzo *et al.* (1994), there is a probability of p that a mutation changes the microsatellite length by 1 unit and has a probability of $1 - p$ that a change in length is ≥1 repeat unit(s), where the length of change is given by a geometric distribution. The generalized mutational model (Fu and Charkraborty 1998) is a simplified version of this mixed model, which sets p to 0, and consequently the length of change is entirely governed by the geometric distribution.

Many population genetics inference methods for microsatellite data require the adoption of a mutation model such as those described above. These approaches can be divided into three categories. The first category involves moment estimators based on summary statistics, including sample homozygosity (Kimmel *et al.* 1998; Xu and Fu 2004) and allele length variance (Wehrhahn 1975; Kimmel *et al.* 1998), to estimate $\theta = 4N_e\mu$ (four times the product of effective population size and the mutation rate).

The second category consists of likelihood-based approaches to the estimation of θ. As it is not in general possible to evaluate the likelihood function analytically, it is approximated by computational methods including Markov chain Monte Carlo (MCMC) (Beerli and Felsenstein 1999) and importance sampling (Nielsen 1997). Significant progress has been made in the development of methods that employ importance sampling and composite likelihoods for microsatellite inference, allowing the maximum-likelihood estimate of demographic parameters to be computed efficiently (Iorio *et al.* 2005; RoyChoudhury and Stephens 2007). On the other hand, Wilson and Balding (1998), Beaumont (1999), and others have applied MCMC to provide Bayesian inference of demographic history from microsatellite data, in which case population parameters are treated as random variables instead of unknown fixed parameters as in a maximum-likelihood approach. Cornuet *et al.* (2006) investigates the underlying mutation process of microsatellites using reversible-jump MCMC (Green 1995) of microsatellite models.

The third category includes likelihood-free approaches such as approximate Bayesian computation (ABC) (Weiss and Von Haeseler 1998; Beaumont *et al.* 2009; Bertorelle *et al.* 2010). The application of ABC to microsatellite data (Beaumont *et al.* 2002; Cornuet *et al.* 2008; Tallmon *et al.* 2008) aims to increase computation efficiency as the method uses summary statistics instead of the full data set and employs simulation to circumvent the likelihood computation step.

Many of the likelihood approaches mentioned above are based on the coalescent theory (Kingman 1982; Griffiths and Tavare 1994). Rather than assuming a parametric model for the population history, for example exponential growth or logistic growth models (Pybus *et al.* 2003), advanced coalescent-based methods provide inference of the demographic history by estimating population as a function of time directly from the data (Drummond *et al.* 2005; Opgen-Rhein *et al.* 2005; Heled and Drummond 2008; Minin *et al.* 2008), but most of them have not been accessible for microsatellite inference.

To extend previous work on Bayesian coalescent inference of microsatellite data, we develop a method that provides inference of the demographic history averaged over a nested set of microsatellite mutation models that incorporate length dependency, mutation bias, and step size. Our method can handle multiple loci and these are assumed to be unlinked or in independent blocks of linkage. The implementation of this method consists of two main parts. The first part is to introduce the implementation of a rich family of microsatellite mutational models and other necessary components to the BEAST software package (Drummond and Rambaut 2007), which provides microsatellite inference access to sophisticated coalescent models (Drummond *et al.* 2005; Heled and Drummond 2008; Minin *et al.* 2008). The second part deals with model uncertainty by employing the product space formulation of transdimensional MCMC (Sisson 2005) as described in section 2.5 of Godsill (2001), which facilitates Bayesian model selection by producing posterior probabilities of the microsatellite mutation models and Bayesian model averaging for estimates of population history and genealogies over those models. The transdimensional MCMC technique chosen here uses techniques from Bayesian variable selection (BVS) (Geweke 1996; Kuo and Mallick 1998) *sensu* Godsill (2001). The BVS-inspired scheme is preferred over other transdimensional MCMC techniques because it trades a small increase in MCMC state space for a high degree of simplification and flexibility in programming.

To apply BVS a composite model space must be constructed that nests all submodels of interest over which inference of population history and genealogies should be averaged. In our case the submodels have a natural nesting by variable selection, because each model represents a special case of the most general microsatellite mutation model in our family. However, models that do not nest naturally can still be averaged over using BVS by introducing a simple index parameter alongside the union of all submodel parameters to construct a product space over models (Carlin and Chib 1995). However, an advantage of our nested model space is that it is able to indicate which of the three microsatellite mutation properties has a strong signal in the data.

MATERIALS AND METHODS

The basic global model: Here we give an overview of the global model and the framework within which our implementation is developed. The microsatellite data, $\mathbf{D}$, consist of L loci, $\mathbf{D} = \{\mathbf{D}_1, \ldots, \mathbf{D_L}\}$ and each locus is composed of a collection of microsatellite repeat lengths from the population of interest, $\mathbf{D}_l = \{\mathbf{D}_{l1}, \ldots, \mathbf{D}_{ln_l}\}$, where $l = 1, \ldots, L$ and n_l is the number of copies of locus l collected. In haploid data, n_l is the number of sampled individuals, while in diploid data n_l is twice the number of individuals from which the samples have been collected. We assume that the data have been generated by an underlying continuous time Markov chain (CTMC), along an unknown genealogy τ_l, which is a rooted bifurcating tree. In the simulations and analyses, the mutation rate is assumed constant along the tree within a locus, *i.e.*, a strict molecular clock rate. The time intervals between successive coalescence events in the genealogy are modeled by the coalescent, which requires a demographic model component containing parameters $\boldsymbol{\Theta}$. The mutation process is defined by the microsatellite mutation model (more details in the Microsatellite models section) with parameters $\boldsymbol{\phi}$. Let $\boldsymbol{\tau} = \{\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_L\}$ and assuming the loci are independent and identically distributed given $(\boldsymbol{\tau}, \boldsymbol{\phi})$, the joint posterior distribution of $\boldsymbol{\tau}$, $\boldsymbol{\phi}$, and $\boldsymbol{\Theta}$ is

$$f_{N,G,\Phi}(\boldsymbol{\Theta}, \boldsymbol{\tau}, \boldsymbol{\phi} \mid \mathbf{D}) \propto \prod_{l=1}^{L} \Pr(\mathbf{D}_l \mid \boldsymbol{\tau}_l, \boldsymbol{\phi}) f_G(\boldsymbol{\tau}_l \mid \boldsymbol{\Theta}) f_N(\boldsymbol{\Theta}) f_\Phi(\boldsymbol{\phi}) \quad (1)$$

(Drummond *et al.* 2002). The tree likelihood of locus l is $\Pr(\mathbf{D}_l | \boldsymbol{\tau}_l, \boldsymbol{\phi})$ and can be evaluated using the peeling/pruning algorithm described by Felsenstein (1981), although we employ augmentation of internal nodes with repeat lengths. The coalescent models come into play by serving as priors for the tree topology and coalescent/divergence times. The form of the coalescent likelihood, $f_G(\boldsymbol{\tau}_l \mid \boldsymbol{\Theta})$, depends on a demographic model specified *a priori* and its parameters $(\boldsymbol{\Theta})$ are jointly estimated. The prior distributions for parameters of the demographic model and mutational model are selected from various standard univariate and multivariate distributions.

Microsatellite models: The models of interest in this study were ones that could be approximated by finite state space continuous-time Markov chains to readily incorporate them into existing software for likelihood calculations on trees. We first need to decide on the coding of the data. Unlike nucleotides or amino acids that have a finite state space, the size of the microsatellite state space is ambiguous, because a universal upper bound for length of microsatellites probably cannot be defined (Kruglyak *et al.* 1998). Yet, according to previous observations, the number of repeats in a microsatellite allele rarely exceeds a few tens (Goldstein and Pollock 1997). In addition, it seems sensible to impose a lower bound on repeat length, above which we can expect the characteristic behavior of microsatellite mutation to occur. In this article, an allele with i repeats is denoted as i. The imposed maximum and minimum lengths of a microsatellite are denoted as $i_{\max}$ and $i_{\min}$, respectively. Therefore there are $s = i_{\max} - i_{\min} + 1$ possible states.

Once boundaries are set, it is easy to define the infinitesimal rate matrix of an ergodic Markov chain. The infinitesimal rate matrix $Q := (q_{i,j})_{i,j=i_{\min},\ldots,i_{\max}}$ is a square matrix wherein each element $q_{i,j}$ specifies the relative instantaneous rate of allele i mutating to allele j, and the shared lower and upper bounds of i and j are $i_{\min}$ and $i_{\max}$, respectively. Given the mutation rate (μ), the Markov chain has the transition probability matrix (P),

$$P(\mu t) := \left(p_{i,j}\right)_{i,j=i_{\min}}^{i_{\max}} = e^{-Q\mu t},$$

where $p_{i,j}$ (μt) is the probability of mutating from allele i to allele j, in time t.

In our implementation, the infinitesimal rate matrix of the most complex model is parameterized:

$$q_{i,j} = \begin{cases} \alpha(u_0, u_1, u_2, d_0, d_1, d_2, i)(1 + (1-p)\gamma(g, i, j)), & |i-j| = 1 \\ \alpha(u_0, u_1, u_2, d_0, d_1, d_2, i)(1-p)\gamma(g, i, j)), & |i-j| > 1 \\ -\sum\limits_{k \neq i} q_{i,k}, & i = j, \end{cases} \tag{2}$$

or

$$q_{i,j} = \begin{cases} \alpha(1, a_1, a_2, 1, a_1, a_2, i)\beta(b_0, b_1, i)(1 + (1-p)\gamma(g, i, j)), & j = i+1 \\ \alpha(1, a_1, a_2, 1, a_1, a_2, i)\beta(b_0, b_1, i)(1-p)\gamma(g, i, j), & j > i+1 \\ \alpha(1, a_1, a_2, 1, a_1, a_2, i)(1 - \beta(b_0, b_1, i))(1 - (1-p)\gamma(g, i, j)), & j = i-1 \\ \alpha(1, a_1, a_2, 1, a_1, a_2, i)(1 - \beta(b_0, b_1, i))(1-p)\gamma(g, i, j), & j < i-1 \\ -\sum\limits_{k \neq i} q_{i,k} & \text{if } i = j. \end{cases} \tag{3}$$

$$\alpha(u_0, u_1, u_2, d_0, d_1, d_2, i) = \begin{cases} u_0 + u_1(i - i_{\min}) + u_2(i - i_{\min})^2 & \text{if } j = i+1 \\ d_0 + d_1(i - i_{\min}) + d_2(1 - i_{\min})^2 & \text{if } j = i-1 \end{cases}$$

$$\beta(b_0, b_1, i) = \frac{1}{1 + e^{-(b_0 + b_1(i - i_{\min}))}},$$

$$\gamma(g, i, j) = \begin{cases} \dfrac{(1-g)g^{|i-j|-1}}{1 - g^{i_{\max} - i}} & \text{if } i_{\min} \leq i < j \leq i_{\max} \\ \dfrac{(1-g)g^{|i-j|-1}}{1 - g^{i - i_{\min}}} & \text{if } i_{\min} \leq j < i \leq i_{\max}. \end{cases}$$

The rate matrix is normalized so that the total mutational outflow is 1.0; *i.e.*, let $q'_{i,i} = cq_{i,i}$ and find c so that $-\sum_i q'_{i,i}\pi_i = 1.0$.

The function $\alpha(u_0, u_1, u_2, d_0, d_1, d_2, i)$ is the truncated version of the asymmetric quadratic model proposed by Calabrese and Durrett (2003) and accounts for the length dependency of mutation rate and mutational bias by modeling the rate of expansion and contraction as separate quadratic equations. The rate can be symmetric if expansion and contraction share exactly the same quadratic equation; in other words $u_0 = d_0$, $u_1 = d_1$, and $u_2 = d_2$. Equal rate for all lengths is obtained by setting the coefficients of the linear terms and quadratic terms to zero in both equations. Similarly, the rates are modeled as linear functions of the length when u_2 and d_2 are set to 0.

In Equation 3 the α-function has symmetric rates. It is standardized so that parameters u_0 and d_0 are fixed to 1.0, because the rate of $i_{\min}$ must be a positive real number and for any

$\alpha(1, a_1, a_2, 1, a_1, a_2, i)$ it is equivalent to constant × $\alpha(1, a_1, a_2, 1, a_1, a_2, i)$, because of the normalization of the rate matrix.

The focal length is equal to i_f if $i_{\min} \le i_f \le i_{\max}$ and is the repeated root of the equation $(u_0 - d_0) + (u_1 - d_1)(i_f - i_{\min}) + (u_2 - d_2)(i_f - i_{\min})^2 = 0$. At the focal length the rate of expansion and contraction is the same, so given a mutation event, there is equal probability of expansion and contraction.

Although the function $\alpha(u_0, u_1, u_2, d_0, d_1, d_2, i)$ has taken mutational bias into account, the parameterization may not necessarily provide answers to questions regarding the relationship between mutational bias and microsatellite length. As mentioned earlier, mutational bias can be quantified by the probability of expansion given a mutation event. The function $\beta(b_0, b_1, i)$ models the probability of expansion by a simple logistic regression.

Both the bias constant parameter, b_0, and the bias linear parameter, b_1, take real values from the range $(-\infty, +\infty)$. The probability of contraction is $1 - \beta(b_0, b_1, i)$. This is a modification of the parameterization adopted by Sainudiin *et al.* (2004). They model expansion probability by simple linear regression. The probability of expansion then becomes

$$\beta'(b'_0, b'_1, i) = \max\{0, \min\{1, b'_0 - b'_1(i - i_{\min})\}\}. \tag{4}$$

In Equation 4 the bias constant parameter is $b'_0 \in [0, 1]$, while the bias linear parameter is $b'_1 \in (-\infty, +\infty)$.

It is worth noting that $\beta(b_0, b_1, i) \in (0, 1)$ whereas $\beta'(b'_0, b'_1, i) \in [0, 1]$. The difference may not seem significant computationally, but $\beta'(b'_0, b'_1, i)$ can give rise to numerical instability, when there are several consecutive rows of zero expansion (or contraction) rates in the infinitesimal rate matrix. This situation is very rare since in reality it is unlikely that probability of expansion will be close to 0 or 1; it can occur, however, if the user does not specify the appropriate starting values for b'_0 and b'_1. In addition, it is conventionally more appropriate to model a categorical variable with logistic regression (Agresti 2002).

Both parameterizations β and β' were implemented. Parameters of a logistic regression are not as easy to interpret as those of a linear regression, so for a more straightforward interpretation, β' can be chosen for the analysis.

To account for larger steps in state space by a single mutation, we employ the parameterization used by Sainudiin *et al.* (2004), which is similar to the TPM proposed by Di Rienzo *et al.* (1994). Under this model, single-repeat mutations have a probability of p whereas multirepeat mutations (length of change ≥1 repeat) have a probability of $1 - p$. For multirepeat mutations, the distribution of step size, $|i - j|$, is given by a truncated geometric distribution $\gamma(g, i, j)$. The symbol g is the failure probability of the truncated geometric distribution.

Stationary distribution: For an ergodic Markov chain, as time, t, approaches positive infinity, its transition probability matrix converges to a matrix in which every single row is the stationary distribution, $\lim_{t\to\infty} P(\mu t) = 1\pi$. As mentioned by Sainudiin *et al.* (2004) the stationary distributions of all one-phase models are special cases of the general birth–death chain.

Bayesian model uncertainty: The output of a Bayesian analysis is the posterior distribution of the parameters given the data. However, the high-dimensional parameter space in a genealogy-based analysis dictates simulation of the posterior distribution by computationally intensive Monte Carlo methods such as MCMC or importance sampling. Here, the posterior distribution is produced by the Metropolis–Hastings MCMC algorithm (Metropolis *et al.* 1953).

In a Bayesian framework, the standard procedure to compare two models is by computing their Bayes factor (BF), which is the ratio of the marginal likelihoods of the two models (M_1 and M_2):

$$\begin{aligned} \mathrm{BF} &= \frac{\int \Pr(\phi_1|M_1)\Pr(D \mid \phi_1, M_1)\,d\phi_1}{\int \Pr(\phi_2|M_2)\Pr(D \mid \phi_2, M_2)\,d\phi_2} \\ &= \frac{\Pr(D|M_1)}{\Pr(D|M_2)} = \frac{\Pr(M_1|D)/\Pr(M_1)}{\Pr(M_2|D)/\Pr(M_2)}. \end{aligned} \tag{5}$$

If the space of potential models is large, then some techniques for model *comparison* are very time consuming. In addition and more importantly, the mutational model may not be of prime interest, *i.e.*, nuisance, and therefore it is not ideal to perform a separate analysis for every mutation model. The solution to this problem is Bayesian model averaging (BMA). We employ transdimensional MCMC to provide joint inference via sampling the microsatellite model indicator, ν, to produce the posterior distribution

$$\begin{aligned} & f_{N,G,\Phi,V}(\boldsymbol{\Theta}, \tau, \phi, v|\boldsymbol{D}) \\ & \propto \prod_{l=1}^{L} \Pr(\mathbf{D}_l \mid \boldsymbol{\tau}_l, \boldsymbol{\phi}, v) f_G(\boldsymbol{\tau}_l \mid \boldsymbol{\Theta}) f_N(\boldsymbol{\Theta}) f_{\Phi,V}(\boldsymbol{\phi}, v). \end{aligned} \tag{6}$$

Here, $\boldsymbol{\phi}$ is a union of parameter vectors of all n models of interest, and $\boldsymbol{\phi} = \cup_{v=1}^{n} \phi_{\mathbf{Mv}}$, where $\phi_{\mathbf{Mv}}$ is the parameter vector of model M_ν. The marginal posterior distribution of model indicator ν can be obtained from posterior samples of $\Pr(\boldsymbol{\Theta}, \boldsymbol{\tau}, \phi, v|D)$, representing the posterior distribution of the microsatellite model. The joint posterior distribution of $\boldsymbol{\Theta}$ and $\boldsymbol{\tau}$ integrated over the models is

$$\begin{aligned} f_{N,G}(\boldsymbol{\Theta}, \boldsymbol{\tau} \mid \boldsymbol{D}) &= \sum_{v \in V} \int f_{N,G,\Phi,V}(\boldsymbol{\Theta}, \boldsymbol{\tau}, \boldsymbol{\phi}, v \mid \boldsymbol{D})\,d\boldsymbol{\phi} \\ &= \sum_{v \in V} f_{N,G,\Phi}(\boldsymbol{\Theta}, \boldsymbol{\tau}|v, \boldsymbol{D}) f_V(v \mid \boldsymbol{D}), \end{aligned} \tag{7}$$

which can also be expressed as the model-averaged posterior joint distribution of $\boldsymbol{\Theta}$ and $\boldsymbol{\tau}$.

Our implementation of transdimensional MCMC combines the techniques of BVS (Kuo and Mallick 1998) and pseudopriors or linking densities (Carlin and Chib 1995). Early BVS applications have aimed to solve the problem of variable selection encountered when building a linear regression model. Initially, there is a large number of potential predictors $\mathbf{X_1}, \ldots, \mathbf{X_p}$, with values x_{ij}, $j = 1, \ldots, p$ and the focus is on determining which of these predictors are linearly associated with the response variable Y. The full model describes the response y_i as a linear combination of the explanatory variables x_{ij}:

$$y_i = \phi_0 + \sum_{j=1}^{p} \phi_j x_{ij} + \varepsilon_i. \tag{8}$$

The term ϕ_0 is the intercept, and the error term $\varepsilon_i \sim N(0, \sigma^2)$. A coefficient ϕ_j that is (statistically) significantly different from 0 suggests predictor $\mathbf{X_j}$ may help in predicting the response. Conversely, a ϕ_j that does not significantly differ from 0 indicates $\mathbf{X_j}$ provides little additional information and can be excluded from the model. The variable selection method by Kuo and Mallick (1998) uses an auxiliary binary indicator variable, $\boldsymbol{\delta}$, of P dimension. $\delta_j = 1$ indicates the presence and $\delta_j = 0$ indicates the absence of the parameter ϕ_j. The full linear model becomes

$$\begin{aligned} y_i &= \phi_0 + \sum_{j=1}^{p} \phi_j \delta_j x_{ij} + \varepsilon_i \\ &= \phi_0 + \sum_{j=1}^{p} \psi_j x_{ij} + \varepsilon_i. \end{aligned} \tag{9}$$

The term ψ_j can be considered as the outcome of the function $\psi_j = g(\phi_j, \delta_j) = \phi_j\delta_j$. Setting δ_j to 0 forces ψ_j to 0, so that ϕ_j is effectively excluded from the model. However, even though in such a case the value of ϕ_j has no effect in the likelihood, it is still sampled by the MCMC machinery, but according to its prior distribution only. This means the dimensions of $\boldsymbol{\phi}$ and $\boldsymbol{\delta}$, and hence the model parameter dimension, are not changed even though mutation model parameters are effectively included and excluded in the likelihood during the MCMC. Therefore it does not require the computation of the Jacobian ratio unlike transdimensional MCMC techniques such as reversible-jump (RJ)MCMC (Green 1995).

We augment the parameters in Equation 3 with a set of indicators, each associated with one of the parameters in the most general microsatellite mutation model, to produce a natural nesting of the described microsatellite mutation models.

Our most complex (full) microsatellite model (defined by Equation 3) contains the parameters, $\boldsymbol{\phi}^{\mathbf{Full}} = \{a_1, a_2, b_0, b_2, g, p\}$, and any submodel has only a subset of $\boldsymbol{\phi}^{\mathbf{Full}} \cdot \phi_k^{\mathrm{Full}}$, $k = \{1, \ldots, 6\}$, is the kth element in $\boldsymbol{\phi}^{\mathbf{Full}}$. We augment $\boldsymbol{\phi}^{\mathbf{Full}}$ with a binary indicator variable $\boldsymbol{\delta} = \{\delta_1, \ldots, \delta_6\}$ to provide a set of toggle switches that can be used to define all nested models of $\boldsymbol{\phi}^{\mathbf{Full}}$. Letting $q_{ij}(\boldsymbol{\phi}^{\mathbf{Full}})$ represent Equation 3, the equation that defines the instantaneous rate matrix of the new model is $q_{ij}^{\mathrm{tdMCMC}} = q_{ij}(\psi)$ and our function for ψ is given by

$$\psi_k = h(\phi_k, \delta_k) = \delta_k \phi_k^{\mathrm{Full}}. \tag{10}$$

Thus the instantaneous rate matrix depends on both $\boldsymbol{\phi}^{\mathbf{Full}}$ and $\boldsymbol{\delta}$. Even when the value of ϕ_k^{Full} has no effect on the likelihood as $\delta_k = 0$, ϕ_k^{Full} is still sampled by the MCMC machinery, but according to its prior distribution only. The joint posterior distribution when using this model is

$$\begin{aligned} &\Pr(\boldsymbol{\Theta}, \boldsymbol{\tau}, \boldsymbol{\phi}, \boldsymbol{\delta} \mid \mathbf{D}) \\ &\quad \propto \prod_{l=1}^{L} \Pr(\mathbf{D}_l \mid \boldsymbol{\tau}, \boldsymbol{\phi}, \boldsymbol{\delta}) f_G(\boldsymbol{\tau} \mid \boldsymbol{\Theta}) f_N(\boldsymbol{\Theta}) f_{\Phi\mid\Delta}(\boldsymbol{\phi} \mid \boldsymbol{\delta}) f_\Delta(\boldsymbol{\delta}), \end{aligned} \tag{11}$$

where $\boldsymbol{\phi} = \boldsymbol{\phi}^{\mathbf{Full}}$. If $\boldsymbol{\phi}$ and $\boldsymbol{\delta}$ are assumed independent, $f_{\Phi\mid\Delta}(\boldsymbol{\phi} \mid \boldsymbol{\delta}) f_\Delta(\boldsymbol{\delta})$ is replaced by $f_\Phi(\boldsymbol{\phi}) f_\Delta(\boldsymbol{\delta})$.

Prior on model space: There are six free parameters in the full model, which theoretically give us 64 submodels. However, parameter p cannot be estimated for submodels in which g is not a free parameter (*i.e.*, fixed to 0). This is because if g is fixed to 0, p does not have any effect on the likelihood. Furthermore, when modeling with regression, it is convention to estimate all polynomial terms in the model up to the largest degree considered in the model. We apply this convention to functions $\alpha(1, a_1, a_2, 1, a_1, a_2)$ and $\beta(b_0, b_1, i)$ in Equation 3. The application of these restrictions to the model space results in a connected subspace of 27 models, and we apply a uniform prior so that the prior probability on each model is 1/27, while the remaining 37 models have a prior probability of 0.0. Figure A1 in appendix a shows the restricted model space.

Proposal distributions: Model switching is performed by two proposal moves, the flip move and the pick move. The flip move uniformly picks an index of the bit vector $\boldsymbol{\delta}$ at random and performs a flip, whereby the value at that index changes from 0 to 1 or vice versa. For a bit vector of n dimension, the probability is $1/n$ for both the flip move and its reverse. The Hastings ratio for $M_i \rightarrow M_j$ is $q(M_i|M_j)/q(M_j|M_i) = (1/n)/(1/n) = 1.0$. This proposal distribution over bit vector $\boldsymbol{\delta}$ is symmetric and therefore no Hastings correction is required for this proposal. Since, in our case, the 27 models with nonzero prior probability form a connected component, the flip move will produce an ergodic and irreducible Markov chain. In effect, the nonhomogeneity of the restricted model space is corrected for by rejection of the neighboring models with zero prior probability, rather than defining a Hastings ratio specifically for the restricted model space (shown in Figure A1 of appendix a).

The pick move, on the other hand, allows larger moves. It selects a model uniformly at random from the set of 27 permitted models. All permitted models have equal probability to be selected, and thus the pick move is symmetric.

Pseudopriors: Pseudopriors or linking densities, a technique used in transdimensional MCMC, were first introduced by Carlin and Chib (1995). Their method considers the situation when there is no overlap among the individual parameter vectors of the n models of interest. The parameter "pool" $\boldsymbol{\phi}$ is therefore simply a concatenation of all model parameter vectors. The joint posterior distribution for v and $\boldsymbol{\phi}$ given data $\mathbf{D}$ can be written as

$$\begin{aligned} \Pr(v, \boldsymbol{\phi} \mid \mathbf{D}) &\propto \Pr(\mathbf{D} \mid v, \boldsymbol{\phi}) f(\boldsymbol{\phi} \mid v) f(v) \\ &\propto \Pr\big(\mathbf{D} \mid v, \boldsymbol{\phi}_{\mathbf{M_v}}\big) f\big(\boldsymbol{\phi}_{\mathbf{M_v}} \mid v\big) f\big(\boldsymbol{\phi}_{-\mathbf{M_v}} \mid v, \boldsymbol{\phi}_{\mathbf{M_v}}\big) f(v), \end{aligned}$$

where the $\boldsymbol{\phi}_{-\mathbf{M_v}} = \bar{\boldsymbol{\phi}}_{\mathbf{M_v}} \cap \boldsymbol{\phi}$, and their values do not affect the likelihood when the current model is M_ν. The expression $f(\boldsymbol{\phi}_{-\mathbf{M_v}} \mid \nu, \boldsymbol{\phi}_{\mathbf{M_v}})$ represents the "pseudopriors" by Carlin and Chib (1995) and can be considered the prior distributions of parameters in $\boldsymbol{\phi}_{-\mathbf{M_v}}$, when (by definition) their values are not being used by the likelihood. If $\boldsymbol{\phi}_{-\mathbf{M_v}}$ is assumed independent of $\boldsymbol{\phi}_{\mathbf{M_v}}$, the pseudopriors become $f(\boldsymbol{\phi}_{-\mathbf{M_v}} \mid v, \boldsymbol{\phi}_{\mathbf{M_v}}) = f(\boldsymbol{\phi}_{-\mathbf{M_v}} | v)$. Unlike "real priors," these pseudopriors have no effect on the joint posterior $f(v, \boldsymbol{\phi}_{\mathbf{M_v}} | Y)$, but govern the mixing of MCMC as they play the role of jumping distributions in RJMCMC (Green 2003). Appropriate pseudopriors resemble efficient proposal distributions and achieve efficient sampling by preventing extremely unlikely parameter values of $\boldsymbol{\phi}_{-\mathbf{M_v}}$ from being sampled.

Selecting suitable pseudopriors can overcome the problem of poor mixing encountered when the prior is very different from the posterior for parameters being model averaged. When a parameter is not in the likelihood, values sampled from the prior may have little agreement with the data. Consequently, a parameter may have difficulty reentering the likelihood, resulting in poor mixing.

Godsill (2001) extended the method by Carlin and Chib (1995) to allow arbitrary overlap among parameter vectors $\boldsymbol{\phi}_{\mathbf{M_v}}$. In addition to a pool of parameters, $\boldsymbol{\phi}$, indicators, $I(\nu)$, map ν to the elements of $\boldsymbol{\phi}$ used by model M_v. The posterior is expressed as

$$\Pr(v, \boldsymbol{\phi} \mid \mathbf{D}) \propto \Pr\big(\mathbf{D} \mid v, \boldsymbol{\phi}_{\mathbf{I(v)}}\big) f\big(\boldsymbol{\phi}_{\mathbf{I(v)}} \mid v\big) f\big(\boldsymbol{\phi}_{-\mathbf{I(v)}} \mid v\big) f(v). \tag{12}$$

Carlin and Chib (1995) suggested that the pseudoprior of a parameter ϕ in model M_ν should match closely the model-specific posterior distributions $\Pr(\phi|Mv)$. It has been observed in some trial runs that even though two models M_{v1} and M_{v2} share the parameter ϕ, the marginal posterior distributions $\Pr(\phi \mid M_{v1})$ and $\Pr(\phi \mid M_{v2})$ are quite different. However, a parameter can have only a single pseudoprior. To achieve model

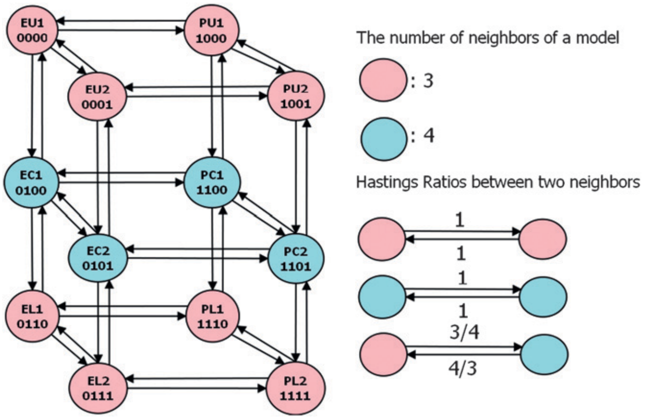

Figure 1.—Restricted model space of the 12 models considered in the simulation analyses.

specificity of pseudopriors, we can augment the parameter space, so that ϕ in M_{v1} is a separate parameter from that in M_{v2}, thus allowing different pseudopriors.

To accommodate augmented mutation model parameter space for a model-specific pseudoprior, the new function for ψ becomes

$$\psi_{\kappa} = \begin{cases} \phi_{I(k,\delta)} & \delta_k = 1 \\ 0 & \delta_k = 0, \end{cases} \qquad (13)$$

where the mapping function $I(k, \delta)$ returns the index of the element in $\boldsymbol{\phi}$ according to k, which indicates the type of parameter, and $\boldsymbol{\delta}$, which specifies the currently active model. For example, if $\boldsymbol{\phi} = \{a_1, a_2, b_0^1, b_0^2, b_1, g, p\}$ and b_0^1 is the constant bias parameter for models of the form $**10**$, whereas b_0^2 is that for $**11**$ models, then $\phi_{I(k=3,\delta=**11**)}$ maps to the parameter b_0^1, where * is either 0 or 1. Again, when the rest of the parameters in the pool are not used in the likelihood, they are sampled from their respective pseudopriors and so the parameter dimension remains the same.

Tree likelihood computation: Felsenstein's (1981) pruning algorithm of tree likelihood computation implicitly sums over all possible ancestral states. For a data type with s states and a rooted gene tree with n taxa ($n - 1$ ancestral nodes), the pruning algorithm is $O(ns^2)$. The speed of this algorithm is sufficiently fast for analysis to be completed on nucleotide data, which have state space size of 4 (A, T, C, and G). For microsatellite DNA, however, the number of states is many times larger than that of the nucleotide data type, and therefore likelihood calculation is much more time consuming.

One solution to this problem is to avoid summation across all possible states at ancestral nodes by treating unknown ancestral allelic states, $\mathbf{D}_l^{\mathbf{A}}$, as auxiliary parameters (Wilson and Balding 1998). After augmentation of the tree with ancestral states and fixing to a particular microsatellite model $\mathbf{M}_{\mathbf{v}}$ with parameters $\boldsymbol{\phi}_{\mathbf{M}_{\mathbf{v}}}$, the tree likelihood of loci l is the product of all likelihoods of nodes in a tree,

$$L^{(l)} = \Pr\left(\mathbf{D}_l, \mathbf{D}_l^{\mathbf{A}} \mid \boldsymbol{\tau}_l, \boldsymbol{\phi}_{\mathbf{M}_{\mathbf{v}}}\right) = \pi_{i_{\text{root}}} \prod_{x=1}^{2n-2} \Pr\left(i_x \mid i_{\text{anc}(x)}, t_x, \boldsymbol{\phi}_{\mathbf{M}_{\mathbf{v}}}\right), \qquad (14)$$

where x is one of the $2n - 2$ nodes in the tree excluding the root and i_x is the state of node x. The parent of node x is denoted as $\text{anc}(x)$ and t_x is the length of the branch that connects x to $\text{anc}(x)$. Following Wilson and Balding (1998), we replace Felsenstein's tree likelihood with the likelihood in Equation 14. The prior probability of the ancestral state in the root node, $\pi_{i_{\text{root}}}$, is computed from the stationary distribution of the mutational process, as is standard in Felsenstein's likelihood of an independent ergodic Markov process on a tree. Ancestral states in the remaining internal nodes have a uniform prior.

For a discussion on the numerical stability see appendix b.

Proposal moves for ancestral state sampling: The candidate allelic state of an ancestral node is proposed by a random-walk integer move, which makes a step from the current allelic state. This move randomly picks direction and step size, which is an integer between 1 and a maximum step size specified by the user. The maximum step size permitted is less than the difference between the upper and lower boundaries of the allelic state. If after a random-walk step the value proposed exceeds the boundaries, then the exceeding proportion of the step is reflected back. Due to the condition on the maximum step size and the type of reflection chosen, the result of a reflection will not be on either boundary. Given maximum step size, w, the number of possible combinations of direction and step size is $2w$. The Hastings ratio is thus the ratio for a move from state i to j and is given by $H(j, i)/H(i, j)$, where $H(i, j) = h_1(i, j) + h_2(i, j) + h_3(i, j)$. The equations $h_1(i, j)$, $h_2(i, j)$, and $h_3(i, j)$ are given below:

$$h_1(i,j) = \begin{cases} 1, & 0 < |i - j| \leq w \\ 0, & \text{otherwise.} \end{cases}$$

$$h_2(i,j) = \begin{cases} 1, & i + w > i_{\max} \text{ and } 2i_{\max} - (i + w) \leq j \text{ and} j \neq i_{\max} \\ 0, & \text{otherwise.} \end{cases}$$

$$h_3(i,j) = \begin{cases} 1, & i - w < i_{\min} \text{ and } 2i_{\min} - (i - w) \geq j \text{ and} j \neq i_{\min} \\ 0, & \text{otherwise.} \end{cases}$$

The proposal mechanism is independent of the currently indicated model in transdimensional MCMC.

Simulations: After developing the implementation for Bayesian microsatellite analysis, it is of interest to obtain some indication of the accuracy and precision of the estimates. We consider only a subset of the 27 models in the restricted model space. This subset is obtained by setting a_2 and p to 0; therefore the most complex model considered here has only four parameters and the bit vector $\boldsymbol{\delta}$ has four dimensions. Because of the restriction that b_1 is a free parameter only when b_0 is a free parameter, this subset has 12 permitted models instead of 16.

These 12 models resemble the set in Sainudiin *et al.* (2004), except we use simple logistic regression to model mutational bias. For convenience, we use their model naming system. This restricted model space of 12 models is illustrated in Figure 1.

Simulated data were generated under the 12 different microsatellite models from the procedure described below:

1. One hundred replicate data sets were generated under each microsatellite model.
2. For each replicate, 30 random coalescent trees were generated, each with 15 individuals assuming a constant population size with $N_e\mu = 2.0$ (where N_e is the effective population size of chromosomes and μ is the mutation rate representing the number of mutations per microsatellite locus per generation).
3. A microsatellite data type was created with minimum length of 1 repeat unit and a maximum of 30 repeat units.
4. For each coalescent tree, a microsatellite site pattern was simulated under the microsatellite model with mutation rate equal to 1.0. All site patterns in a trial were simulated under the same microsatellite model. There were 15 sampled haploid individuals in each site pattern.

Each of the 1200 simulated unlinked 30-locus data sets were analyzed with transdimensional MCMC to demonstrate how well our method identifies the true microsatellite mutational model (M_{true}). We also compared the accuracy and precision in the demographic estimates between model averaging and when the true model was known. Analyses have chain lengths of 70 million steps with transdimensional MCMC and 50 million steps with the true model. Sampled parameters were recorded every 50,000 steps.

It is also of interest to investigate the effect of the number of taxa and that of loci on the precision of θ-estimation using transdimensional MCMC. Data sets were simulated under the PL2 model with different combinations of number of loci and number of taxa presented in supporting information, Table S1, which also includes the MCMC chain length for each combination. One hundred simulations were carried out for each combination. For this set of simulations, we recorded every 10,000th step of the MCMC. The convergence of each analysis was examined by the trace analyses including the computation of the effective sample size (ESS) of each estimated parameter. Table S2 is a summary of the model parameter values that are chosen for simulating the data. All simulated data sets are provided in .csv format in File S1.

Measure of accuracy: The accuracy was measured by computing the relative bias. Here we define relative bias as

$$\mathrm{bias}_k = \frac{\hat{\theta}_{k,\mathrm{median}} - \theta}{\theta},$$

where θ is the true population size value and $\hat{\theta}_{k,\mathrm{median}}$ is the posterior median population size for trial k.

Accuracy of estimates may also be indicated from the percentage of trials with 95% highest probability intervals (95% HPD) containing the true answer. The α% HPD is the smallest interval containing α% of the posterior distribution.

Measure of precision: The relative error was used to measure the precision of the estimates obtained. We define the relative error as

$$\mathrm{error}_k = \frac{|\hat{\theta}_{k,\mathrm{median}} - \theta|}{\theta}.$$

Another measure of precision is the 95% HPD relative bound and is given as

$$95\%\ \text{HPD relative bound} = \frac{95\%\ \text{HPD upper bound} - 95\%\ \text{HPD lower bound}}{\theta}.$$

The credible interval coverage, relative bias, error, and HPD bound defined here are similar to the corresponding measures used by Heled and Drummond (2008).

Prior distribution for microsatellite model parameters: We used a normal(0, 10) prior on both b_0 and b_1, an exponential(1) on a_1, and uniform(0,1) on g.

Pseudopriors: From test runs it appears that only b_0, b_1, and g require pseudopriors to reach reasonable convergence. The pseudopriors for each variable are chosen to be tight distributions centered around the true parameter values since they were known. Pseudopriors for each parameter are shown in Table S3.

Tree prior: For the simulations, the tree prior used was the coalescent with constant population (see Kingman 1982 or Griffiths and Tavare 1994 for details on the coalescent likelihood calculation). For inference, the constant population size model was chosen to match the simulation conditions.

The prior density for θ was set to one-on-x prior, $f(x) \propto 1/x$. The one-on-x prior is an improper prior; however, in the case of constant population size, it can be shown to be Jeffrey's prior, and its application in this context leads to a proper posterior distribution (Drummond *et al.* 2002, 2004).

Microsatellite model prior: Because we did not have any *a priori* information regarding the microsatellite model, a uniform prior is applied to the set of 12 models considered.

Sampling tree topologies: The tree proposal moves subtree-slide, narrow exchange, wide exchange, and Wilson–Balding are used for tree topology sampling in all the analyses undertaken in this article. A nice summary of these proposal moves is presented in Höhna *et al.* (2008).

Red colobus monkey data: *Data:* The red colobus monkey (*Pilocolobus tephrosceles*) data set was kindly provided by J. Allen (University of Florida, Gainesville). This unpublished data set consists of 62 samples from each of 16 loci. Each locus was typed for both homologous copies from each of 31 red colobus monkeys from the Kibale National Park of Uganda. These loci are treated as unlinked as no clear signal of linkage had been found (J. Allen, personal communication). The allele lengths and the PCR primers are presented in Table S4 and Table S5.

Analyses: An upper bound of 33 and lower bound of 6 repeats were imposed. We made boundaries wider than the observed length range of the data to account for the possibility that the observed sample range is not the true range in the population. We ran two separate analyses of 200,000,000 for each of the 12 microsatellite models. We also ran two replicate analyses using transdimensional MCMC. To compare mixing and performance between transdimensional MCMC and fixing the microsatellite model, we estimated θ from this data set with transdimensional MCMC and each of the 12 models used for simulation. Values for ESS per MCMC step were computed for θ, tree likelihoods, coalescent likelihoods, and mutation rates. We accommodated the potential mutation rate variation across loci by estimating the relative rates but fixing the average rate to 1.0.

Prior selection: A uniform Dirichlet prior for 16 dimensions was applied to the relative mutation rates. The tree prior and mutation model parameter priors used for the real data analyses are the same as the ones for analyses of the simulation data. However, because the true values of the mutational parameters are unknown, pseudopriors could not be picked as easily as for simulated data. We took Carlin and Chib's suggestion and obtained preliminary posterior distributions of mutation model parameters by running a short MCMC (of 40,000,000 steps) with each microsatellite model. The posterior densities obtained from these preliminary runs are still quite broad, but they provide sufficient guidance for the selection of pseudopriors. The first 10% of each chain is discarded as burn-in and the remaining chain is used to fit a standard parametric distribution to the posterior sample of each parameter.

The marginal posterior distributions of the microsatellite mutation parameters were fitted using the maximum-likelihood–based fitdistr function in the MASS package (Venables and Ripley 2002) of R (R Development Core Team 2009), a software environment for statistical computing. The fitdistr function returns parameter estimates for a parametric distribution that best describes the posterior sample. The quality of the fit is then examined by the one-sample Kolmogorov–Smirnov test with the null hypothesis that the posterior sample has come from the fitted distribution. Several different parametric distributions were fitted to the sample and the one with the largest Kolmogorov–Smirnov test P-value (least evidence against a bad fit) was chosen.

TABLE 1

Percentage of true model recovery computed from transdimensional MCMC (tdMCMC) analysis of simulated data

	M_{Best}											
M_{True}	EU1	EU2	EC1	EC2	EL1	EL2	PU1	PU2	PC1	PC2	PL1	PL2
EU1	*95*	4	0	0	0	0	1	0	0	0	0	0
EU2	13	*87*	0	0	0	0	0	0	0	0	0	0
EC1	0	0	*59*	0	1	0	0	0	37	3	0	0
EC2	2	5	13	*63*	0	1	0	0	2	15	0	0
EL1	18	0	0	0	*59*	6	1	0	0	0	17	0
EL2	0	0	0	0	10	*76*	0	0	0	0	6	14
PU1	0	0	0	0	0	0	*97*	3	0	0	0	0
PU2	0	0	0	0	0	0	17	*83*	0	0	0	0
PC1	0	0	21	0	0	0	0	0	*73*	6	0	0
PC2	0	1	1	5	0	0	0	0	18	*75*	0	0
PL1	4	0	0	0	4	1	3	0	0	0	*78*	10
PL2	0	0	0	0	0	1	8	5	0	0	21	*65*

Table S6 presents all the pseudoprior distributions. In this case, the parameter space is augmented so that there is no overlap in model parameter vectors and each parameter has its own model-specific pseudoprior.

RESULTS

Simulations: For each analysis of simulated data, the first 10% of the chain was discarded as burn-in and analysis of traces confirm that all parameters have ESS > 100. The accuracy of Bayesian model selection by our method is indicated by how often the maximum *a posteriori* model corresponds with the true model, M_{True}. Table 1 presents the frequency distribution of M_{Best} (the model with maximum posterior probability using our transdimensional MCMC method) for data sets simulated under each M_{True}. The highest percentage value in each row is in italics. It is shown that all diagonal values are in italics, which means $M_{\text{Best}} = M_{\text{True}}$ has the highest frequency for all 12 models of M_{True}.

Accuracy is also indicated by computing the percentage of trials that has the M_{True} contained in the 95% HPD set of models. These values are presented in Table 2. A very high proportion (>0.9) of the trials captures M_{True} within the 95% HPD set. The median 95% HPD set size is between two and four models, and the majority of the models have a median set size of two, suggesting good precision.

Often, the user is more concerned with the accuracy of other evolutionary or demographic estimates rather than the mutational model *per se*. The values for relative bias, relative error, and relative 95% HPD bounds for the demographic parameter of constant population size are calculated for each trial. Table 3 is a summary of the median relative bias, the median relative error, the median 95% HPD relative bound, and the percentage of trials in which the 95% HPD interval captured the true value of $N_e\mu = 2.0$ for each model.

Estimates with high precision have small values of median relative error or median 95% HPD relative bound. Accurate estimates have small values of median relative bias and a high percentage of 95% HPD intervals containing the true value. Within each row, the values of the four statistics of accuracy and precision are close between BMA and when the true model is known. However, when the true model (TM) is known, the results have greater coverage of the true population parameter value, smaller median relative error, smaller median relative 95% HPD bound, and smaller absolute median relative bias than model-averaged estimates of θ.

All model–method combinations had high frequentist coverage varying from 0.86 to 0.99. Given the small number of replicates, these coverage statistics are not significantly different from each other and are all consistent with an underlying proportion in the region of 0.95, although in the Bayesian setting there is no reason to expect coverage to be at the 0.95 level.

For either method, there is a spectrum of median relative error values across the models of M_{True}, where the median relative error value is the smallest when the

TABLE 2

Accuracy and precision of true model recovery

M_{True}	Inside 95% HPD	95% HPD set size
EU1	1.00	2
EU2	1.00	2
EC1	1.00	3
EC2	0.98	3
EL1	0.95	4
EL2	0.99	2
PU1	1.00	2
PU2	1.00	2
PC1	1.00	3
PC2	0.99	2
PL1	0.98	3
PL2	0.93	2

TABLE 3

Measure of accuracy and precision of model-averaged θ-estimates from transdimensional analyses and of θ-estimates from analyses that fixed the microsatellite mutational model to the true model

	Inside 95% HPD		Median relative error		Median relative bias		Median relative bound	
M_{True}	BMA	TM	BMA	TM	BMA	TM	BMA	TM
EU1	0.97	0.98	0.10	0.10	−0.06	−0.02	0.58	0.55
EU2	0.94	0.97	0.12	0.10	0.07	0.04	0.83	0.73
EC1	0.96	0.93	0.12	0.12	0.02	0.01	0.68	0.61
EC2	0.89	0.91	0.21	0.14	0.09	0.11	0.97	0.77
EL1	0.89	0.92	0.10	0.10	−0.02	0.01	0.59	0.55
EL2	0.92	0.92	0.16	0.12	0.14	0.08	0.95	0.70
PU1	0.98	0.99	0.10	0.09	−0.06	−0.03	0.65	0.62
PU2	0.94	0.99	0.22	0.18	0.15	0.10	1.01	0.87
PC1	0.93	0.93	0.11	0.10	−0.01	0.03	0.65	0.63
PC2	0.86	0.92	0.16	0.14	0.10	0.08	0.95	0.87
PL1	0.90	0.93	0.11	0.12	−0.04	0.03	0.32	0.31
PL2	0.89	0.95	0.15	0.13	0.07	0.00	0.80	0.75

data are simulated under PU1 and largest when data are simulated under PU2.

Similarly, there is quite some variation in the size of the median relative 95% HPD bounds across different models of M_{True} for both BMA and analysis with the true model. Median relative 95% HPD bounds are the smallest for data simulated under PL1 and the largest for PU2, which are approximately three times of that for PL1.

Precision vs. number of loci: Values of median relative error and median relative HPD range are plotted against the number of loci (Figure 2), where the number of taxa is fixed to 15. As the number of loci increases, the relative error (Figure 2, dashed blue line) and 95% HPD range (Figure 2, solid red line) reduce linearly (on a log-log scale). It appears that increasing the number of loci to eight times larger will halve the relative error and reduces the 95% HPD range by a factor >2.

Precision vs. number of taxa: Figure 3 shows median relative error and 95% HPD credible intervals as a function of the number of taxa per loci for 10 unlinked loci. Both the error (Figure 3, dashed blue line) and the HPD interval (Figure 3, solid red line) decrease as the number of taxa increases, but they seem to asymptote to some positive limits. Further reduction of error and HPD range can be achieved only by sampling more loci. The function form $y = a_0 + a_1/(a_2 + x)$ is chosen merely to illustrate the general trend.

Real data example—colobus monkey data: Tracer (Rambaut and Drummond 2007) was used to decide the length of the chain to be discarded from a log file as burn-in. The burn-in length was ∼10–20% of the original length of each log file. The two logs from analyses with the same microsatellite model were subsequently combined. The combined log files were then examined again by Tracer to investigate mixing and convergence to stationarity. All ESS values were >150.

According to the results from transdimensional MCMC, the model with the highest posterior probability is EU1. The 95% highest posterior probability set consists of EU1, EU2, PL1, EL1, and EL2, with the respective probabilities 0.483, 0.324, 0.053, 0.049, and 0.048. The posterior probability of including a parameter $a_1 = 0.089$, $b_0 = 0.193$, $b_1 = 0.186$, and $g = 0.412$. These posterior probabilities suggest that multistep mutation is the most evident feature followed by mutation bias and rate dependency.

The posterior mean, median, and 95% HPD interval from analyses with each model are presented in Table 4.

The posterior median of θ for single-step models ranges from 4.28 to 4.97 and that for multistep models ranges from 3.40 to 3.98. Since a multistep model was sampled almost half the time, the model-averaged posterior median of θ is somewhere in between.

Mixing and performance: We use ESS values per MCMC step as an indication of the sampling efficiency. Figure S1, Figure S2, Figure S3, Figure S4, and Figure S5 are dot plots of ESS value per MCMC step for θ, tree likelihoods, coalescent likelihoods, root heights, and mutation rates. These plots suggest that the sampling efficiency of transdimensional MCMC is only slightly less than the average sampling efficiency of single-model analyses. The ratio of ESS per MCMC step for transdimensional MCMC *vs.* single-model analyses on average is 0.78 for θ. Averaging across loci, the ratio is 0.90 for tree likelihood, 0.74 for coalescent likelihood, 0.71 for root height, and 0.98 for relative mutation rate.

DISCUSSION

The focus of this research was on the implementation of a nested family of microsatellite mutation models in the BEAST software package (Drummond and Rambaut 2007). There are many analysis tools unique to BEAST, including nonparametric coalescent-based inference methods such as the extended Bayesian skyline plot (Heled and Drummond 2008) and the newly

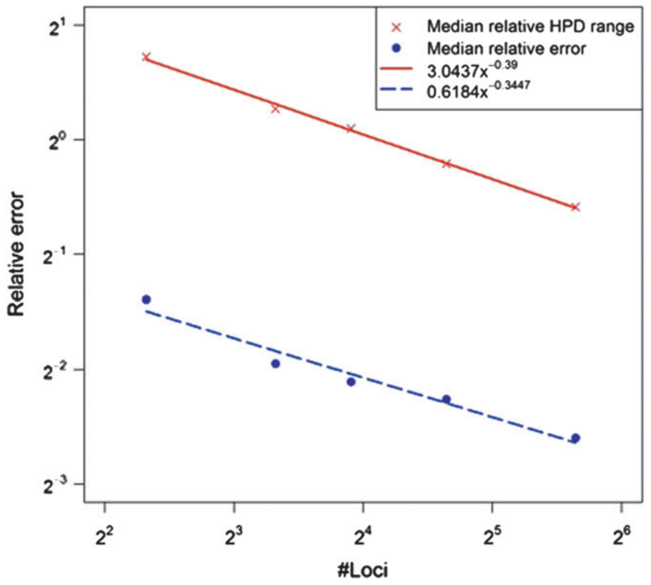

Figure 2.—Measures of precision θ-estimation *vs.* the number of loci.

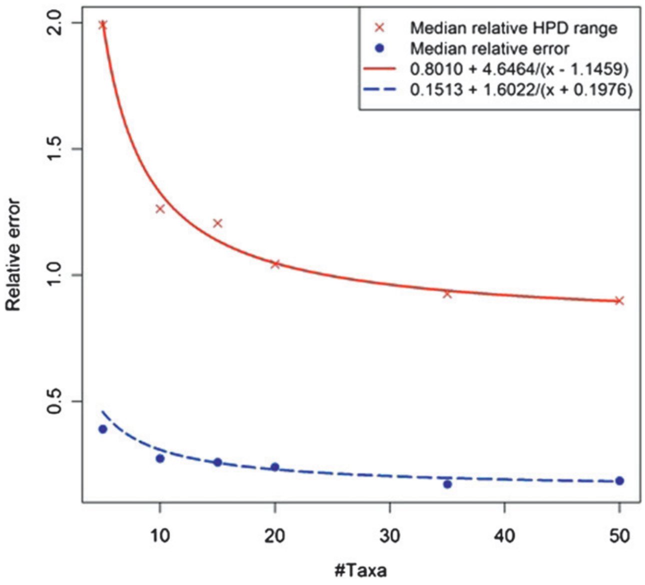

Figure 3.—Measures of precision θ-estimation *vs.* the number of taxa.

developed multispecies coalescent method *BEAST (Heled and Drummond 2010). Equipping BEAST with microsatellite models and other related software components permits the application of these methods to microsatellite data.

The microsatellite models implemented have all been previously described in the literature (Di Rienzo *et al.* 1994; Fu and Charkraborty 1998; Calabrese and Durrett 2003; Sainudiin *et al.* 2004). It was not intended to introduce new models of microsatellite evolution in this work, except to make slight modifications where it made the models more suitable for Bayesian inference in an MCMC setting. Aside from their implementation, simulations were used to investigate their statistical properties.

Simulations: Simulation results show moderate variation in performance across data sets simulated under different microsatellite models. However, when a smaller number of loci were simulated (10), more biased results were observed (results not shown).

As mentioned in parameter prior specification (see section on *Prior distribution for microsatellite model parameters*), free parameters shared by more than one microsatellite model have been given the same prior distribution for all models containing the parameter. For example, the constant bias parameter, b_0, is in models EC1, EC2, PC1, and PC2, and the prior distribution for b_0 is the same for all four models. The impact of the prior choices made for the mutational parameters has not been investigated in this study. It is quite possible that different prior choices would have altered the statistical properties of the estimators. For parameters such as g in the two-phase models, which are defined on [0, 1], the uniform prior is natural; however, for other scale parameters, a number of alternatives are feasible. Therefore, unsuitable prior distributions on the microsatellite model parameters may be partially responsible for the demographic estimation bias observed.

Red colobus monkey data example: Our results suggest that the convergence speed of transdimensional MCMC is only slightly worse than for single microsatellite model analyses on average. In addition, only one analysis of transdimensional MCMC is required to perform model selection; however, for single-model analyses, we would need as many as the number of models of interest (12 in this case) and thus require a far longer time.

In this analysis we have selected pseudoprior densities by fitting univariate distributions to densities acquired from preliminary runs. The procedure can become much less time consuming if an empirical density function is used, coupled with the automation of the input file preparation for preliminary runs and the transdimensional MCMC. However, poorly mixed preliminary runs may still yield pseudopriors that are very different from the posterior and thus offer little improvement in mixing of the analysis with transdimensional MCMC. If human data are analyzed, then appropriate pseudopriors may also be constructed from the wealth of empirical data on the mutation parameters from sperm-typing (Zhang *et al.* 1994) and pedigree studies (Weber and Wong 1993; Xu *et al.* 2000; Whittaker *et al.* 2003).

Model averaging and Bayes factors: In addition to providing a model-averaged posterior distribution of population and genealogical parameters, our transdimensional MCMC method also facilitates robust estimates of the marginal likelihoods (and therefore Bayes factors) of the individual microsatellite models. These estimates are not subject to the large, even infinite variances (Raftery *et al.* 2007; Calderheada and Girolami 2009) associated with the harmonic mean estimator of

TABLE 4

Estimates of θ from red colobus monkey data

Model	Mean	Median	95% HPD lower	95% HPD upper
BMA	4.08	4.01	2.56	5.76
EU1	4.37	4.31	3.02	5.83
EU2	3.54	3.47	2.25	4.98
EC1	4.33	4.28	3.06	5.75
EC2	3.46	3.40	2.27	4.83
EL1	4.62	4.54	3.26	6.15
EL2	3.87	3.80	2.56	5.38
PU1	4.33	4.28	2.93	5.92
PU2	3.50	3.45	2.11	4.94
PC1	4.60	4.48	2.85	6.68
PC2	3.59	3.49	2.06	5.28
PL1	5.05	4.97	3.33	6.88
PL2	4.05	3.98	2.45	5.72

marginal likelihoods (Newton and Raftery 1994). Using transdimensional MCMC to estimate Bayes factors is also computationally efficient as it requires only a single MCMC run to determine the relative merits of all k mutational models, rather than the k (or more) independent runs required by other techniques, including thermodynamic integration (Lartillot and Philippe 2006).

Issues and improvement: *Low information content in a single microsatellite locus:* Low information content of a single microsatellite locus means inference results may be sensitive to poor prior choices. In comparison to microsatellite data, mtDNA sequence data possess much more information available for reconstruction of the genetic ancestry. As a result, an mtDNA tree has a higher level of resolution than a microsatellite tree. However, this does not mean inference on mtDNA sequence data is more accurate than that on microsatellite data. Given a population history, the coalescent admits wide variation in the topology and coalescent times of the gene trees. To make a more reliable inference, it is important to use multiple loci, each of which has an independent history (Felsenstein 2006). Even though mtDNA sequence data provide a clear view of the genealogy of the mtDNA sequences, the whole mtDNA genome is a completely linked locus. The mtDNA tree therefore provides only one of the many possible realizations of the coalescent process for a given population history. On the contrary, there are potentially thousands of independent microsatellite loci available to overcome the problem of stochastic variability of individual genealogies.

Speed and convergence: A large microsatellite data set with hundreds of loci may give very accurate population size estimates, but is currently not practical in our implementation, due to slow convergence and computational inefficiencies. An analysis for a data set containing ~60 unlinked loci and 100 taxa requires days to satisfy our heuristic diagnostic statistics for convergence when all the loci were used simultaneously in the same MCMC run. The slow convergence is due to the large joint parameter space when all loci are unlinked. The parameter space containing 60 independent 100-tip trees is much larger than that having a single 100-tip tree with 60 linked sites. One potential solution is sequential Monte Carlo methods (Liu 2001), which take advantage of the independence structure of the likelihood to build up a full posterior distribution by sequential analysis of the loci (De Finetti 1974). Besides the speed, there is also the need to improve the efficiency of sampling ancestral states. Our implementation samples ancestral states by a naive Metropolis–Hastings algorithm and therefore has low acceptance probability. Gibbs sampling (Geman and Geman 1984) is an alternative MCMC algorithm that can sometimes produce more efficient sampling. It is a special case of the Metropolis–Hastings algorithm, whereby each proposed candidate is always accepted, since the components of the state that change in the proposal are drawn directly from the conditional posterior distribution. Gibbs sampling of internal nodes may improve the convergence.

Comparison with other software: Well-known software programs such as BATWING (Wilson *et al.* 2003) and Migrate (Beerli 2004) also provide Bayesian coalescent analysis on microsatellite data. However, these programs contain only a few simple microsatellite models. BATWING provides the SMM and the *K*-allele model; microsatellite model options in Migrate are the SMM (called the ladder model in Migrate) and Brownian motion (an approximation of the SMM). These models do not take into account many properties of microsatellite mutation and as mentioned in the Introduction there is much evidence for those properties. Therefore the simplifying assumptions of the SMM may not be adequate to perform inference on real data. Furthermore, we provide model averaging over a rich set of microsatellite models, which is absent in these programs.

In the case of BATWING, all microsatellites are assumed to be linked into a single locus. It was discussed earlier that incorporation of multiple loci is necessary for accurate inference. Using only a single locus overlooks the genome-wide distribution of microsatellites, a highly advantageous trait for coalescent inference.

Future directions: In all analyses in this study, all loci shared the same (model-averaged) microsatellite model within an MCMC run. It is possible that the properties of mutation vary across loci. While our implementation allows for variation in both rates and microsatellite models across loci, we have not performed a systematic study of the properties of such models. A hierarchical prior structure can account for variation of the same component in different models. For example, every locus could have its own EC1 model, containing the parameter b_0. During the MCMC, each b_0 varies according to a given distribution, and the parameters of this distribution (hyperparameters) also have a prior.

Our framework for analysis of microsatellite data can be combined with the multispecies coalescent (Heled and Drummond 2010) to estimate the species tree using multiple microsatellite loci sampled from closely related species. Although microsatellite models can be used alongside various relaxed-clock models in BEAST (Drummond *et al.* 2006) to estimate divergence times, we do not recommend this type of analysis, because each microsatellite locus does not have sufficient information to estimate rate heterogeneity among branches.

Model selection has always been an important problem in statistical inference. It is common to make inferences on the basis of the best model selected by a standard model comparison procedure. However, such a procedure may produce a subset of models that are not significantly different from one another in their goodness-of-fit and therefore create difficulty in deciding which model provides the most reliable inference. Our transdimensional method allows the data to speak for themselves and more importantly makes population inference on the basis of a set of microsatellite models, accounting for model uncertainty and avoiding model misspecification.

The computer simulations in this article were performed using computational resources provided by BestGRID (http://www.bestgrid.org/), a New Zealand not-for-profit organization that delivers services and tools supporting research. We thank Raazesh Sainudiin and three anonymous reviewers for helpful comments. We thank J. Allen for the unpublished red colobus monkey data set. The collection of the red colobus monkey data set was funded by the National Science Foundation under a grant to D. L. Reed (DEB 0717165). The authors were supported by Marsden grant UOA0809.

LITERATURE CITED

Agresti, A., 2002 *Categorical Data Analysis*, Ed. 2. John Wiley & Sons, New York.

Beaumont, M., J. Cornuet, J. Marin and C. Robert, 2009 Adaptive approximate Bayesian computation. Biometrika **96:** 983–990.

Beaumont, M., W. Zhang and D. Balding, 2002 Approximate Bayesian computation in population genetics. Genetics **162:** 2025–2035.

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

Beerli, P., 2004 Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. Mol. Ecol. **13:** 827–836.

Beerli, P., and J. Felsenstein, 1999 Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics **152:** 763–773.

Bertorelle, G., A. Benazzo and S. Mona, 2010 ABC as a flexible framework to estimate demography over space and time: some cons, many pros. Mol. Ecol. **19:** 2609–2625.

Calabrese, P. P., and R. T. Durrett, 2003 Dinucleotide repeats in the Drosophila and human genomes have complex length-dependent mutation processes. Mol. Biol. Evol. **20:** 715–725.

Calabrese, P. P., R. T. Durrett and C. F. Aquadro, 2001 Dynamics of microsatellite divergence and proportional slippage/point mutation models. Mol. Biol. Evol. **159:** 839–852.

Calderheada, B., and M. Girolami, 2009 Estimating Bayes factors via thermodynamic integration and population MCMC. Comput. Stat. Data Anal. **53:** 4028–4045.

Carlin, B. R., and S. Chib, 1995 Bayesian model choice via Markov chain Monte Carlo methods. J. R. Stat. Soc. Ser. B Methodol. **57:** 473–484.

Cornuet, J., M. Beaumont, A. Estoup and M. Solignac, 2006 Inference on microsatellite mutation processes in the invasive mite, Varroa destructor, using reversible jump Markov chain Monte Carlo. Theor. Popul. Biol. **69:** 129–144.

Cornuet, J., F. Santos, M. A. Beaumont, C. P. Robert, J. Marin *et al.*, 2008 Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. Bioinformatics **24:** 2713–2719.

de Finetti, B., 1974 *Theory of Probability*. John Wiley & Sons, New York.

Di Rienzo, A., A. C. Peterson, J. C. Garza, A. M. Valdes, M. Slatkin *et al.*, 1994 Mutational process of simple-sequence repeat loci in human populations. Proc. Natl. Acad. Sci. USA **91:** 3166–3170.

Drummond, A., G. Nicholls, A. Rodrigo and W. Solomon, 2004 Genealogies from time-stamped sequence data, pp. 149–171 in *Tools for Constructing Chronologies: Crossing Disciplinary Boundaries*. Springer-Verlag, Berlin/Heidelberg, Germany/New York.

Drummond, A. J., and A. Rambaut, 2007 BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. **7:** 214.

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. J., A. Rambaut, B. Shapiro and O. Pybus, 2005 Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. **22:** 1185–1192.

Drummond, A. J., S. Y. W. Ho, M. J. Phillips and A. Rambaut, 2006 Relaxed phylogenetics and dating with confidence. PLoS Biol. **4:** 699–710.

Felsenstein, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17:** 368–376.

Felsenstein, J., 2006 Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences or more loci? Mol. Biol. Evol. **23:** 691–700.

Fu, Y., and R. Charkraborty, 1998 Simultaneous estimtion of all the parameters of a step-wise mutation model. Genetics **150:** 487–497.

Geman, S., and D. Geman, 1984 Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Patt. Anal. Mach. Intell. **6:** 721–741.

Gentle, J., 2007 *Matrix Algebra: Theory, Computations, and Applications in Statistics*. Springer-Verlag, New York.

Geweke, J., 1996 Variable selection and model comparison in regression. Bayesian Stat. **5:** 609–620.

Godsill, S. J., 2001 On the relationship between Markov chain Monte Carlo methods for model uncertainty. J. Comput. Graph. Stat. **10:** 230–248.

Goldstein, D. B., and A. G. Clark, 1995 Microsatellite variation in North American populations of Drosophila melanogaster. Nucleic Acids Res. **23:** 3882–3886.

Goldstein, D. B., and D. D. Pollock, 1997 Launching microsatellites: a review of mutation processes and methods of phylogenetic inference. J. Hered. **88:** 335–342.

Green, P., 2003 Trans-dimensional Markov chain Monte Carlo, pp. 179–198 in *Highly Structured Stochastic System*, edited by P. Green N. Hjort and S. Richardson. Oxford University Press, London/New York/Oxford.

Green, P. J., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82:** 711–732.

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

Heled, J., and A. J. Drummond, 2008 Bayesian inference of population size history from multiple loci. BMC Evol. Biol. **8:** 289.

Heled, J., and A. Drummond, 2010 Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. **27:** 570–580.

Höhna, S., M. Defoin-Platel and A. Drummond, 2008 Clock-constrained tree proposal operators in Bayesian phylogenetic inference, pp. 1–7 in *Eighth IEEE International Conference on BioInformatics and BioEngineering, 2008*. IEEE, Athens, Greece.

Iorio, M. D., R. C. Griffiths, R. Leblois and F. Rousset, 2005 Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. Theor. Popul. Biol. **68:** 41–53.

Jackson, C. H., 2011 Multi-state models for panel data: the msm package for R. Journal of Statistical Software. **38**(8): 1–29.

Kimmel, M., R. Chakraborty, J. P. King, M. Bamshad, W. S. Watkins *et al.*, 1998 Signatures of population expansion in microsatellite repeat data. Genetics **148:** 1921–1930.

Kingman, J. F. C., 1982 The coalescent. Stoch. Proc. Appl. **13**(3): 235–248.

Kruglyak, S., R. Durrett, M. D. Schug and C. F. Aquadro, 1998 Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc. Natl. Acad. Sci. USA **95:** 10774–10778.

Kuo, L., and B. Mallick, 1998 Variable selection for regression models. Sankhyā. Ind. J. Stat. **60:** 65–81.

Lartillot, N., and H. Philippe, 2006 Computing Bayes factors using thermodynamic integration. Syst. Biol. **55:** 195.

Levinson, G., and G. A. Gutman, 1987 High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. Nucleic Acids Res. **15:** 5323–5338.

Liu, J. S., 2001 *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, New York.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller, 1953 Equation of state calculation by fast computer machines. J. Chem. Phys. **21:** 1087–1092.

Metzgar, D., L. Liu, C. Hansen, K. Dybvig and C. Wills, 2002 Domain-level differences in microsatellite distribution and content result from different relative rates of insertion and deletion mutations. Genome Res. **12:** 408–413.

Minin, V. M., E. W. Bloomquist and M. A. Suchard, 2008 Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol. Biol. Evol. **25:** 1459–1471.

Newton, M. A., and A. E. Raftery, 1994 Approximate Bayesian inference with the weighted likelihood bootstrap. J. R. Stat. Soc. Ser. B Methodol. **56:** 3–48.

Nielsen, R., 1997 A likelihood approach to populations samples of microsatellite alleles. Genetics **146:** 711–716.

Ohta, T., and M. Kimura, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genetics **22:** 201–204.

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., 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.

R Development Core Team, 2009 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Raftery, A. E., M. Newton, P. Satagopan and J. Krivitsky, 2007 Estimating the integrated likelihood via posterior simulation using harmonic mean identity. Bayesian Stat. **8:** 1–45.

Rambaut, A., and A. Drummond, 2007 *Tracer v1.4*. http://tree.bio.ed.ac.uk/software/tracer/.

Richard, G. F., and F. Pâques, 2000 Mini- and microsatellite expansions: the recombination connection. EMBO Rep. **1:** 122–126.

RoyChoudhury, A., and M. Stephens, 2007 Fast and accurate estimation of the population-scaled mutation rate, theta, from microsatellite genotype data. Genetics **176:** 1363–1366.

Rubinsztein, D. C., B. Amos and G. Cooper, 1999 Microsatellite and trinucleotide-repeat evolution: evidence for mutational bias and different rates of evolution in different lineages. Philos. Trans. R. Soc. B Biol. Sci. **354:** 1095–1099.

Sainudiin, R., R. T. Durrett, C. F. Aquadro and R. Nielsen, 2004 Microsatellite mutation models: insights from a comparison of humans and chimpanzees. Genetics **168:** 383–395.

Schlötterer, C., R. Ritter, B. Harr and G. Brem, 1998 High mutation rates of a long microsatellite allele in Drosophila melanogaster provide evidence for allele specific mutation rates. Mol. Biol. Evol. **15:** 1269–1274.

Shikano, T., Y. Shimada, G. Herczeg and J. Merlä, 2010 History vs. habitat type: explaining the genetic structure of European nine-spined stickleback (Pungitius pungitius) populations. Mol. Ecol. **19:** 1147–1161.

Sibly, R. M., J. C. Whittaker and M. Talbot, 2001 A maximum-likelihood approach to fitting equilibrium models of microsatellite evolution. Mol. Biol. Evol. **18:** 413–417.

Sisson, S., 2005 Transdimensional Markov chains. J. Am. Stat. Assoc. **100:** 1077–1089.

Slatkin, M., 1995 A measure of population subdivision based on microsatellite allele frequencies. Genetics **139:** 457–462.

Smith, G. P., 1976 Evolution of repeated DNA sequences by unequal crossover. Science **191:** 528–535.

Spong, G., M. Johansson and M. Björklund, 2010 High genetic variation in leopards indicates large and long-term stable effective population size. Mol. Ecol. **9:** 1773–1782.

Streisinger, G., and J. Owen, 1985 Mechanisms of spontaneous and induced frameshift mutation in bacteriophage T4. Genetics **109:** 633–659.

Tallmon, D. A., A. Koyuk, G. Luikart and M. A. Beaumont, 2008 onesamp: a program to estimate effective population size using approximate Bayesian computation. Mol. Ecol. Res. **8:** 299–301.

Venables, W. N., and B. D. Ripley, 2002 *Modern Applied Statistics with S*, Ed. 4. Springer-Verlag, New York.

Walsh, J. B., 1987 Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. Genetics **115:** 553–567.

Weber, J., and C. Wong, 1993 Mutation of human short tandem repeats. Hum. Mol. Genet. **2:** 1123.

Wehrhahn, C. F., 1975 The evolution of selectively similar electrophoretically detectable alleles in finite natural populations. Genetics **80:** 375–394.

Weiss, G., and A. Von Haeseler, 1998 Inference of population history using a likelihood approach. Genetics **149:** 1539.

Whittaker, J., R. Harbord, N. Boxall, I. Mackay, G. Dawson *et al.*, 2003 Likelihood-based estimation of microsatellite mutation rates. Genetics **164:** 781.

Wierdl, M., M. Dominska and T. D. Petes, 1997 Microsatellite instability in yeast: dependence on the length of the microsatellite. Genetics **146:** 768–779.

Wilson, I. J., and D. J. Balding, 1998 Genealogical inference from microsatellite data. Genetics **150:** 499–510.

Wilson, I. J., M. E. Weale and D. J. Balding, 2003 Inference from DNA data: population histories, evolutionary processes and forensic match probabilities. J. R. Stat. Soc. Ser. A **166:** 155–188.

Xu, H., and Y. Fu, 2004 Estimating effective population size or mutation rate with microsatellites. Genetics **166:** 555–563.

Xu, X., M. Peng, Z. Fang and X. Xu, 2000 The direction of microsatellite mutations is dependent upon allele length. Nat. Genet. **24:** 396–399.

Zhang, L., E. Leeflang, J. Yu and N. Arnheim, 1994 Studying human mutations by sperm typing: instability of CAG trinucleotide repeats in the human androgen receptor gene. Nat. Genet. **7:** 531–535.

Communicating editor: M. K. Uyenoyama

APPENDIX A: RESTRICTED MODEL SPACE

Figure A1 represents the restricted model space. The nodes represent the models, each labeled with its bit vector representation, and the arrow-edges are the Hastings ratios of a move from one model to it neighbor. Two models are neighbors if they have only a single difference. The nodes (models) are color coded according to the number of neighbors they have.

In this case the restricted model space is one connected component. However, some prior specification with zero probabilities on a subset of models may result in two or more disjointed components. In such cases the flip move alone cannot produce an ergodic Markov chain; therefore the pick move must be used so that all the models in the restricted space can be proposed.

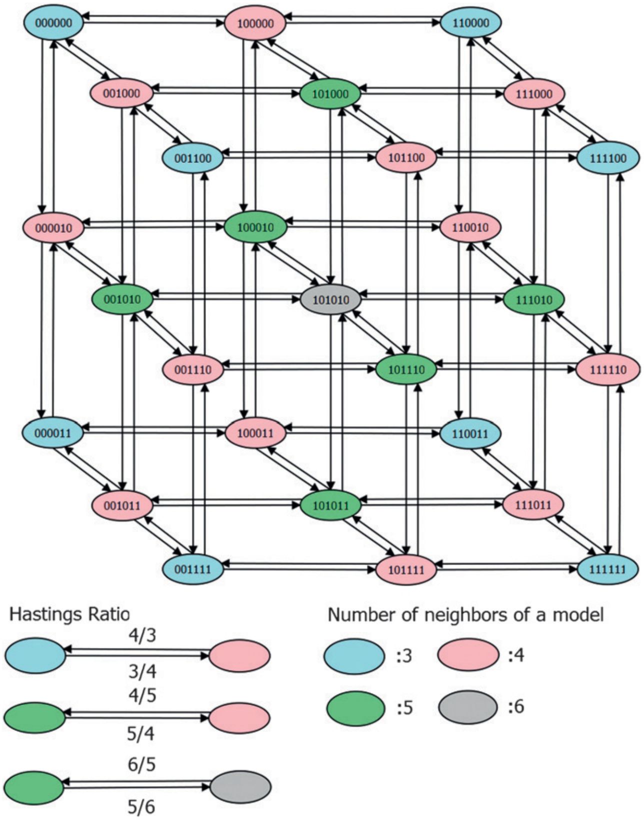

Figure A1.—Reduced model space produced by the restrictions described in the *Prior on model space* section in materials and methods.

APPENDIX B: NUMERICAL STABILITY

We used a few shortcuts in the tree likelihood calculation. If a proposal move results only in a change of likelihood on a few branches, then we subtract the initial logarithm of partial tree likelihood from the logarithm of full tree likelihood and add the new log partial tree likelihood. This is more efficient than computing the entire full log likelihood.

We have found that if only partial tree likelihood calculation is used, the difference in likelihood between partial and entire likelihood calculation at each step increases as the MCMC proceeds. However, a step that requires entire likelihood calculation will set this difference back to 0. Luckily, in a real analysis, there are many other parameters that will force the calculation of the entire full likelihood rather frequently. We have also provided the option so that the user can force the entire full likelihood computation every n number of likelihood computations.

Another component that is relevant to the numerical stability of the method is matrix exponentiation. Matrix exponentiation is achieved by using codes adapted from Cern Colt library 1.2 (http://acs.lbl.gov/software/colt/, more details in appendix c). To ensure that matrix exponentiation of the CERN colt library is reliable, we compare the matrix exponentiation results computed from our codes with from the function MatrixExp in the msm package (Jackson 2009) of R.

Comparisons were made for five values of t (0.001, 0.01, 0.1, 1, and 10) and 12 different Q matrices, which are the instantaneous matrices under which our data were simulated. There was very little difference between the two exponential methods (all differences $<10^{-13}$).

APPENDIX C: MATRIX EXPONENTIATION

The method of matrix exponentiation here requires the Q matrix to be diagonalizable. This requirement is checked by using the method described in Gentle (2007). If the method indicates that the requirement is not met, a 0 probability matrix is returned, which leads to the rejection of the proposal move. The code adapted from Cern COLT library 1.2 performs an eigen decomposition of the instantaneous rate matrix, so the matrix Q can be expressed as $Q = VUV^{-1}$, where V is a matrix of eigenvectors and the U is a diagonal matrix of eigenvalues. Given some value of t, the exponential of Qt is then obtained by finding the matrix product $e^{Qt} = Ve^{Ut}V^{-1}$. The expression e^{Ut} is also a diagonal matrix, with diagonal values $e^{u_{ii}t}$, where u_{ii} represents the ith diagonal value.

GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.110.125260/DC1

Joint Inference of Microsatellite Mutation Models, Population History and Genealogies Using Transdimensional Markov Chain Monte Carlo

Chieh-Hsi Wu and Alexei J. Drummond

DOI: 10.1534/genetics.110.125260

File S1

Data Sets

File S1 is available for download as a compressed folder (.zip) at http://www.genetics.org/cgi/content/full/genetics.110.125260/DC1.

The name of each folder indicates the microsatellite model under which the data sets in the folder have been simulated and the number of loci and taxa in each data set.

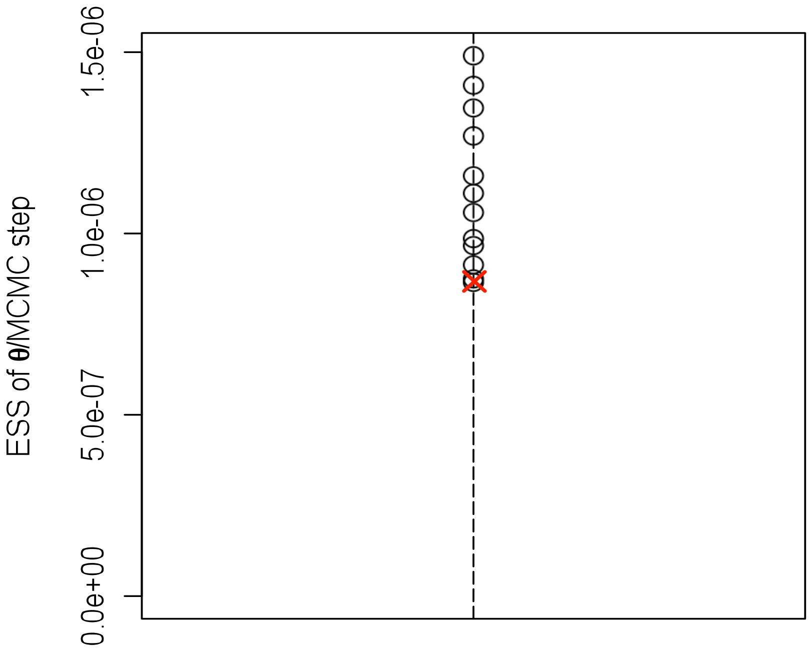

FIGURE S1.—Dot-plot of ESS values of θ per MCMC step. Circles represents analyses with one of the 12 microsatellite models considered, while red cross represents analysis with trans-dimensional MCMC.

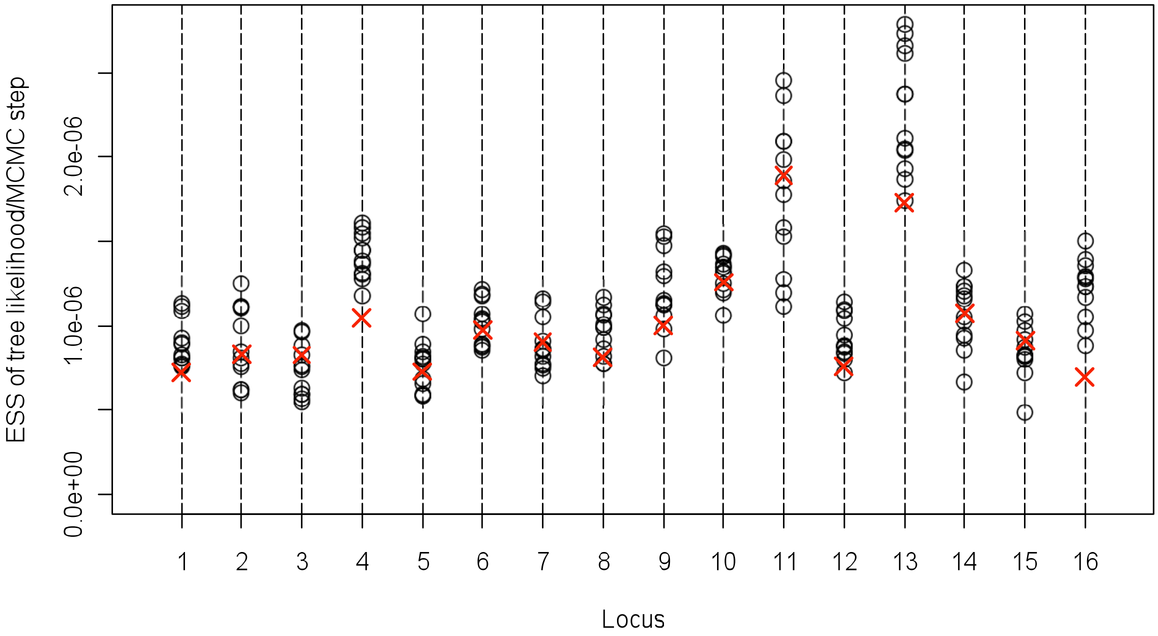

FIGURE S2.—Dot-plot of ESS values of tree likelihoods per MCMC step. Circles represents analyses with one of the 12 microsatellite models considered, while red crosses represents analyses with trans-dimensional MCMC.

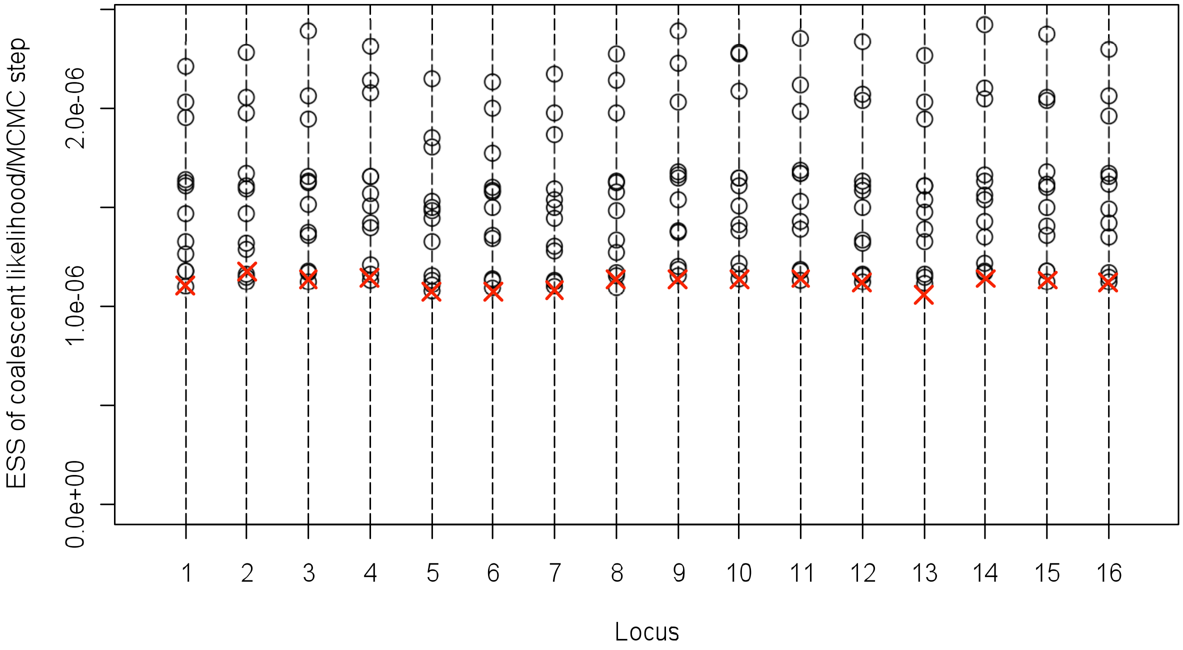

FIGURE S3.—Dot-plot of ESS values of coalescent likelihoods per MCMC step. Circles represents analyses with one of the 12 microsatellite models considered, while red crosses represents analyses with trans-dimensional MCMC.

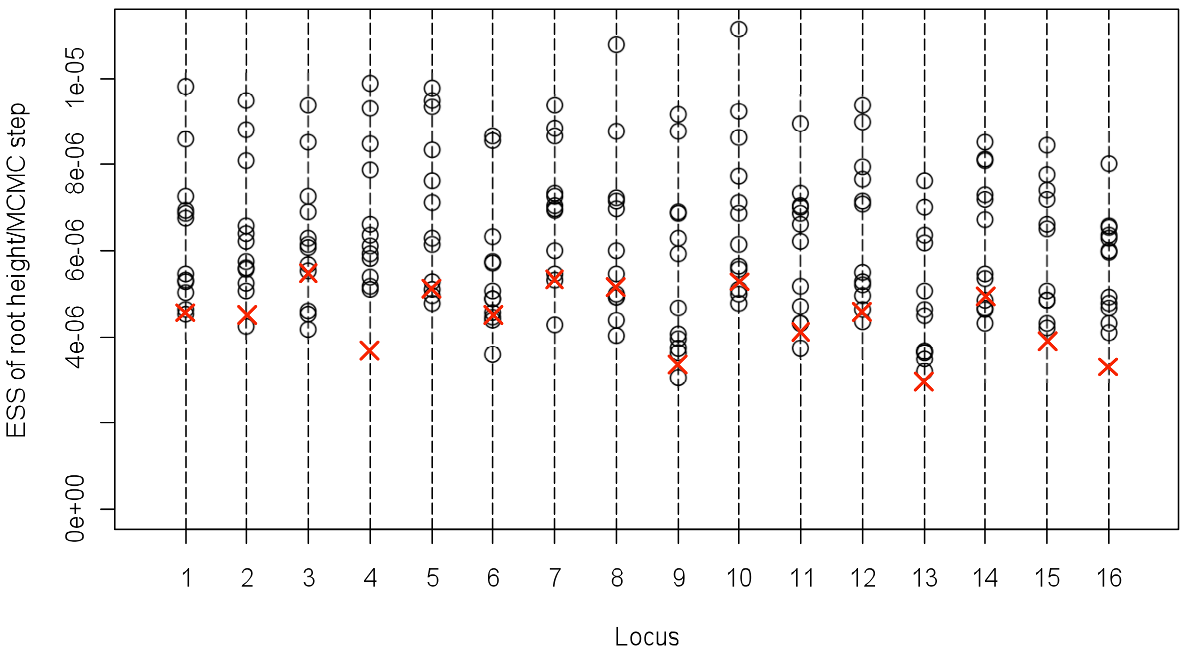

FIGURE S4.—Dot-plot of ESS values of tree root height per MCMC step. Circles represents analyses with one of the 12 microsatellite models considered, while red crosses represents analyses with trans-dimensional MCMC.

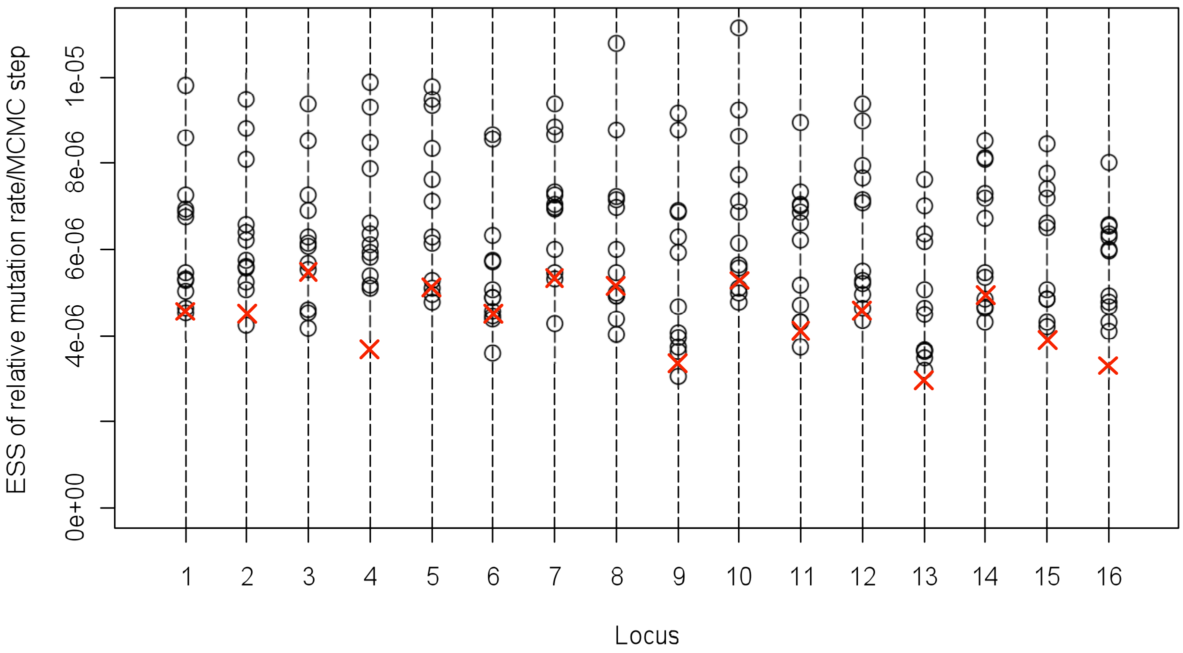

FIGURE S5.—Dot-plot of ESS values of relative mutation rate per MCMC step. Circles represents analyses with one of the 12 microsatellite models considered, while red crosses represents analyses with trans-dimensional MCMC.

TABLE S1

Combination of the number of loci, number of taxa and MCMC chain length for simulations analyses to investigate the effect of the number of loci and taxa on precision of θ estimation

No. of Loci	No. of Taxa	Chain length (10^6 steps)
5	15	10
10	5	10
10	10	10
10	15	20
10	35	50
10	50	100
15	15	30
25	15	60
50	15	80

TABLE S2

Microsatellite model parameter values for simulation

		Parameter values[b]			
	δ[a]	a_1	b_0	b_1	g
M_{True}	[a_1, b_0, b_1,g]	[0, +∞)	(–∞, +∞)	(–∞, +∞)	[0, 1]
EU1	0000	**0.00**	**0.00**	**0.00**	**1.00**
EU2	0001	**0.00**	**0.00**	**0.00**	0.25
EC1	0100	**0.00**	0.20	**0.00**	**1.00**
EC2	0101	**0.00**	0.20	**0.00**	0.25
EL1	0110	**0.00**	0.70	–0.05	**1.00**
EL2	0111	**0.00**	0.70	–0.05	0.25
PU1	1000	0.50	**0.00**	**0.00**	**1.00**
PU2	1001	0.50	**0.00**	**0.00**	0.25
PC1	1100	0.50	0.20	**0.00**	**1.00**
PC2	1101	0.50	0.20	**0.00**	0.25
PL1	1110	0.50	0.70	–0.05	**1.00**
PL2	1111	0.50	0.70	–0.05	0.25

[a] The binary vector **δ** is a the binary representation of the model. The value at each index represents the presence (1) or absence of (0) of a parameter. The rate proportion parameter (a_1) is indicated by the first value of **δ**, the constant bias parameter (b_0) by the second, the linear bias parameter (b_1) by the third and multistep parameter (g) by the fourth.

[b] Values in bold are fixed parameters and therefore fixed during the MCMC run

TABLE S3

Pseudo priors on each parameter in the analysis of simulated data set generated under M_{True}

M_{True}[a]	a_1	b_0	b'_0[b]	b_1	g
EU1	Exp(1)	N(0.0, 0.1)	-	N(0.0, 0.025)	N(0.25, 0.05)
EU2	Exp(1)	N(0.0, 0.1)	-	N(0.0, 0.025)	N(0.25, 0.05)
EC1	Exp(1)	N(0.2, 0.1)	-	N(0.0, 0.025)	N(0.25, 0.05)
EC2	Exp(1)	N(0.2, 0.1)	-	N(0.0, 0.025)	N(0.25, 0.05)
EL1	Exp(1)	N(0.0, 0.1)	N(0.7, 0.1)	N(-0.05, 0.025)	N(0.25, 0.05)
EL2	Exp(1)	N(0.0, 0.1)	N(1.0, 0.1)	N(-0.08, 0.025)	N(0.25, 0.05)
PU1	Exp(1)	N(0.0, 0.1)	-	N(0.0, 0.025)	N(0.25, 0.05)
PU2	Exp(1)	N(0.0, 0.1)	-	N(0.0, 0.025)	N(0.25, 0.05)
PC1	Exp(1)	N(0.2, 0.1)	-	N(0.0, 0.025)	N(0.25, 0.05)
PC2	Exp(1)	N(0.2, 0.1)	-	N(0.0, 0.025)	N(0.25, 0.05)
PL1	Exp(1)	N(0.0, 0.1)	N(0.7, 0.1)	N(-0.05, 0.025)	N(0.25, 0.05)
PL2	Exp(1)	N(0.0, 0.1)	N(1.0, 0.1)	N(-0.08, 0.025)	N(0.25, 0.05)

[a] The model under which the simulated data was generated.

[b] In the analyses of simulated data generated under EL1, EL2, PL1 or PL2, the microsatellite model parameter space has been augmented so that the constant bias parameter (b_0) in EC1, EC2, PC1 and PC2, is treated as a separate parameter to that (b'_0) in *EL1*, *EL2*, *PL1* and *PL2*. Since we did not do the augmentation for analyses of data generated under other models, so the b'_0 is not in the parameter pool in those analyses and the pseudo-prior of b'_0 is therefore not applicable (-).

TABLE S4

Microsatellite allele frequencies from the red colobus monkey population from Kibale National Park in Uganda

Length	Locus															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
8	4	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
9	4	0	0	0	0	0	0	0	45	0	0	0	0	0	0	0
10	18	0	0	25	0	0	0	0	6	0	36	0	0	0	0	0
11	32	0	0	36	4	0	1	0	0	0	18	0	0	0	0	0
12	4	0	0	1	9	0	12	0	9	0	1	2	0	0	4	41
13	0	0	11	0	8	5	10	0	0	0	0	7	0	0	54	10
14	0	0	41	0	17	7	25	0	0	0	1	7	0	0	2	9
15	0	0	10	0	15	11	9	0	0	0	0	26	0	0	2	1
16	0	13	0	0	6	0	2	8	0	0	0	20	0	0	0	0
17	0	17	0	0	3	10	3	9	0	0	0	0	13	0	0	1
18	0	3	0	0	0	11	0	5	0	0	0	0	1	0	0	0
19	0	25	0	0	0	6	0	15	0	0	0	0	0	0	0	0
20	0	4	0	0	0	1	0	22	0	0	0	0	0	0	0	0
21	0	0	0	0	0	11	0	1	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0
24	0	0	0	0	0	0	0	0	0	38	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	24	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	1	15	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	19	31	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0

This data set has been prepared and provided by Dr. J. Allen (University of Florida, USA). Missing data are ignored in the analysis and are not presented.

TABLE S5

PCR primers for the red colobus monkey data

	Forward	Reverse
1	TGAGACCCTGTCTCTGAAAC	TGTATGGGCTCTTGAAATTG
2	AAAGCTACATCCAAATTAGGTAGG	TGACAAAGAAACTAAAATGTCCC
3	TCTGAGCACTCTGGATTGTAGC	ATCTCTGCACGCTTCACTTCTT
4	TACCAACATGTTCATTGTAGATAGA	CATACACCTGTGGACCCATC
5	ACCACATGAGCCAATTCTGT	ACCCAATTATGGTGTTGTTACC
6	CATTGGTCCAGGTAAACTGC	TTCACAAGGTTCCACAAGGT
7	CAAATTAATGGCAAAAACTGC	CCCCCCATTGAGGTTATTAC
8	TCCATTATTCCCCTCAAACA	GGTTTGCCATTCAGTTGAGA
9	AGGCTTGCCAGATAAGGTTG	GCTGAAGGCTGTTCTATGGA
10	ACAAGAGCACATTTAGTCAG	AGCTTCATTTTTCCCTCTAG
11	GTATGATTTATTTCAGGTTTGCA	TTTGATTTCATTGTCTACTGACA
12	TAGGTTCTGGGCATGTCTGT	TGCTTGGCACACTTCAGG
13	CACTTCTCCTTGAATCGCTT	GCAAGTCCTGTTCCAAGTCT
14	ATGCCCTCTTCTGTCTCTCC	GCAGAGAATCTGGACATGCT
15	GCCAACAGAGCAAGACTGTC	GGAAACAGTTAAATGGCCAA
16	GAGAATGTGCCACTGTACTCCA	ACTGGCTCTGAAACTCACCAAT

TABLE S6

Pseudo priors on each parameter after complete augmentation[a] of the microsatellite model parameter space

Model	a_1	b_0	b_1	g
EU1	-	-	-	-
EU2	-	-	-	TN(0.892,0.059)[e]
EC1	-	N(-0.035, 0.022)[b]	-	-
EC2	-	N(-0.045, 0.031)	-	TN (0.873,0.061)
EL1	-	N(0.489, 0.137)	N(0.032,0.007)	-
EL2	-	N(0.548, 0.150)	N(-0.035, 0.008)	TN (0.902, 0.053)
PU1	G(1.376, 15.162)[c]	-	-	-
PU2	LN(2.602, 1.043)[d]	-	-	TN(0.900,0.060)
PC1	LN(2.897,1.335)	N(0.005, 0.028)	-	-
PC2	LN(2.311,1.200)	N(0.009,0.037)	-	TN(0.891,0.053)
PL1	LN(-0.129,0.934)	N(0.637, 0.120)	N(- 0.0367, 0.006)	-
PL2	LN(0.132,1.009)	N(0.704,0.141)	N(-0.040, 0.008)	TN(0.900,0.052)

[a] The model parameter space was augmented such that there was no overlap in parameter vectors among all 12 models considered here. This gave us 24 parameters in the parameter pool and each of them had its own pseudo-prior, assuming they were all independent.

[b] N(μ,σ) is a Normal distribution with mean = μ and standard deviation = σ.

[c] G(α,β) is a Gamma distribution with shape = α and rate = β.

[d] LN(μ,σ) is a Log-normal distribution with log space mean = μ and log space standard deviation = σ.

[e] TN(μ,σ) is a Truncated-Normal distribution which only supports values between 0 and 1 here and have location = μ and scale = σ.