Bayes Estimation of Species Divergence Times and Ancestral Population Sizes Using DNA Sequences From Multiple Loci

Bruce Rannala* and Ziheng Yang[†,1]

*Department of Medical Genetics, University of Alberta, Edmonton, Alberta T6G 2H7, Canada and [†]Galton Laboratory, Department of Biology, University College London, London WC1E 6BT, England

Manuscript received December 4, 2002
Accepted for publication April 18, 2003

ABSTRACT

The effective population sizes of ancestral as well as modern species are important parameters in models of population genetics and human evolution. The commonly used method for estimating ancestral population sizes, based on counting mismatches between the species tree and the inferred gene trees, is highly biased as it ignores uncertainties in gene tree reconstruction. In this article, we develop a Bayes method for simultaneous estimation of the species divergence times and current and ancestral population sizes. The method uses DNA sequence data from multiple loci and extracts information about conflicts among gene tree topologies and coalescent times to estimate ancestral population sizes. The topology of the species tree is assumed known. A Markov chain Monte Carlo algorithm is implemented to integrate over uncertain gene trees and branch lengths (or coalescence times) at each locus as well as species divergence times. The method can handle any species tree and allows different numbers of sequences at different loci. We apply the method to published noncoding DNA sequences from the human and the great apes. There are strong correlations between posterior estimates of speciation times and ancestral population sizes. With the use of an informative prior for the human-chimpanzee divergence date, the population size of the common ancestor of the two species is estimated to be ~20,000, with a 95% credibility interval (8000, 40,000). Our estimates, however, are affected by model assumptions as well as data quality. We suggest that reliable estimates have yet to await more data and more realistic models.

THE (effective) population size N is a central parameter in models of population genetics, conservation genetics, and human evolution. For example, the amount of genetic variation in a population is determined by $\theta = 4N\mu$, where μ is the mutation rate per site per generation. When an independent estimate of the mutation rate is available, we can use the estimate of θ to infer the population size N. Estimation of θ or N of a modern species is relatively simple. The population size of modern humans has been estimated to be ~10,000 (Takahata *et al.* 1995; Ruvolo 1997; Edwards and Beerli 2000; Zhao *et al.* 2000; Yu *et al.* 2001). Estimation of population sizes of extinct ancestors of closely related species is more challenging, but Takahata *et al.* (1995) have developed a maximum-likelihood method under the infinite-sites model for either two or three species. Another commonly used method, for three species, exploits the fact that ancestral polymorphism creates conflicts between the species tree and the gene tree (Nei 1987; Wu 1991) and estimates the ancestral population size by equating the proportion of mismatched gene trees to the theoretical expectation. Application of this method to hominid data sets has led to large estimates of population sizes for the common ancestor of humans and chimpanzees, on the order of 100,000 (Ruvolo 1997; Chen and Li 2001). However, this *tree-mismatch* method ignores sampling errors in the reconstructed gene tree, due to a finite number of nucleotide sites at each locus, and produces serious overestimates (Yang 2002).

Yang (2002) implemented a finite-sites model using both likelihood and Bayes methodologies. The method is limited to the case of three species, with one individual sequenced at each locus from each species. However, it is advantageous to analyze multiple species and loci simultaneously, which may circumvent the sensitivity of such analysis to possible variation in evolutionary rates among loci (Yang 1997; Chen and Li 2001). The realization that typical data do contain information about ancestral population sizes and that information is better extracted by a combined analysis of sequence data from multiple species and loci provided motivation for the present study. Here we extend the method of Yang (2002) to deal with an arbitrary species tree and different numbers of sequences at different loci. The likelihood calculation using numerical integration does not seem feasible due to the increased dimension of the integral. Thus we adopt the Bayes approach and implement a Markov chain Monte Carlo (MCMC) algorithm. We apply the method to published data of noncoding DNA sequences from the human and the great apes.

Computation-intensive MCMC algorithms are in-

[1] *Corresponding author:* Department of Biology, Darwin Bldg., Gower St., London WC1E 6BT, England. E-mail: z.yang@ucl.ac.uk

creasingly being used in inference in molecular population genetics (see Felsenstein *et al.* 1999; Stephens and Donnelly 2000 for reviews). Bahlo and Griffiths (2000) developed a likelihood approach to analyzing sequence data from subdivided populations, estimating jointly the population sizes and migration rates. They used an importance-sampling strategy that works efficiently under the infinite-sites model (Griffiths and Tavaré 1994). Beerli and Felsenstein (1999, 2001) implemented MCMC algorithms for likelihood analysis of subdivided populations under the finite-sites model. Those methods assume an equilibrium migration model, which is suitable for geographically structured populations, but not for the species phylogeny considered in this study, as different species diverged at different times and never reached equilibrium. Nielsen and Wakeley (2001) implemented an MCMC algorithm for likelihood and Bayes inference using data from two species, modeling both ancestral polymorphism and gene flow after the species divergence. More recently Wilson *et al.* (2003) extended the Bayes MCMC algorithm developed for microsatellite data by Wilson and Balding (1998) to account for population subdivision and growth. Wilson *et al.*'s (2003) population-split model allows different subpopulations to diverge at different times, although the total population size is fixed. Their method was not implemented to analyze sequence data. The implementation we present here does not yet account for population demographic processes (such as population growth) or possible gene flow after species divergences, although it is straightforward to add these features. Our method is unique among current methods in being applicable to any species phylogeny and allowing combined analyses of multiple sequences per species as well as sequences from multiple loci.

THEORY

Data and model parameters: The data consist of aligned homologous DNA sequences at multiple neutral loci sampled from present-day species. The model and implementation apply to any species tree. As an example, we focus on the case of the great apes: human (H), chimpanzee (C), gorilla (G), and orangutan (O). The topology of the species tree, (((HC)G)O), is assumed known and fixed in the analysis (Figure 1). The number of sequences sampled may differ among loci. Let $D = \{D_i\}$ be the entire data set, where D_i represent the sequence alignment at locus i, with $i = 1, 2, \ldots, L$ for a total of L loci. We expect the method to be applied to closely related species only and assume the molecular clock, that is, rate constancy among lineages. Furthermore, we assume random mating in each population and no gene flow after species divergences. We also assume no recombination within a locus and free recombination between loci.

Parameters in the model include the species divergence times as well as the ancestral and current population sizes. The population size of a current species is considered only if more than one individual is sampled from that species at some loci. Because time and rate are confounded in the data, both divergence times and population sizes are multiplied by the mutation rate. Thus parameters in the model for the example of Figure 1 include the three divergence times τ_{HC}, τ_{HCG}, and τ_{HCGO} and population size parameters θ_H for humans; θ_C for chimpanzees; and θ_{HC}, θ_{HCG}, and θ_{HCGO} for the three ancestral species. The divergence times (τ's) are measured by the expected number of mutations per site from the ancestral node in the species tree to the present time (Figure 1). Collectively we let $\Theta = \{\theta_H, \theta_C, \theta_{HC}, \theta_{HCG}, \theta_{HCGO}, \tau_{HC}, \tau_{HCG}, \tau_{HCGO}\}$ denote all parameters in the model to be estimated.

Bayes estimation of parameters: The Bayes hierarchical model has two main components: the prior distribution of the parameters and the likelihood, *i.e.*, the probability of the data given the parameters. We use independent gamma distributions as priors for θ's. The gamma density is

$$g(x; \alpha, \beta) = \beta^{\alpha} e^{-\beta x} x^{\alpha-1}/\Gamma(\alpha), \tag{1}$$

with mean α/β and variance α/β^2. The hyperparameters α and β are chosen by the user to reflect the range and likely values of the parameters.

For parameters τ's, we used independent gamma priors for the time gaps (interarrival times) on the species tree. For example, for the species tree of Figure 1, $(\tau_{HCGO} - \tau_{HCG})$, $(\tau_{HCG} - \tau_{HC})$, and τ_{HC} are assumed to have independent gamma distributions, and the prior $f(\Theta)$ is a product of the independent gamma densities. We also implemented an option of specifying gamma priors for the node ages (τ_{HCGO}, τ_{HCG}, and τ_{HC} for the tree of Figure 1), but the prior means are not given by α/β anymore; because of the constraints on node ages (for example, $\tau_{HCGO} > \tau_{HCG} > \tau_{HC}$), the joint gamma distribution is truncated. The MCMC always updates the node ages (τ's) and not time gaps.

The gene genealogy G_i at each locus i is represented by the tree topology T_i and the coalescent times $\mathbf{t}_i$. Given parameters Θ, the probability distribution of $G_i = \{T_i, \mathbf{t}_i\}$ is specified by the coalescent processes under the model. This is described in the next section. Let $G = \{G_i\}$. We have

$$f(G|\Theta) = \prod_i f(G_i|\Theta) = \prod_i f(T_i, \mathbf{t}_i|\Theta). \tag{2}$$

The probability of data D_i given the gene tree and coalescent times (and thus branch lengths) at the locus, $f(D_i|G_i)$, is the traditional likelihood in molecular phylogenetics and can be calculated using any Markov model of nucleotide substitution (Felsenstein 1981). Here we use the model of Jukes and Cantor (1969) to correct for multiple hits at the same site. As we assume independent evolution across loci,

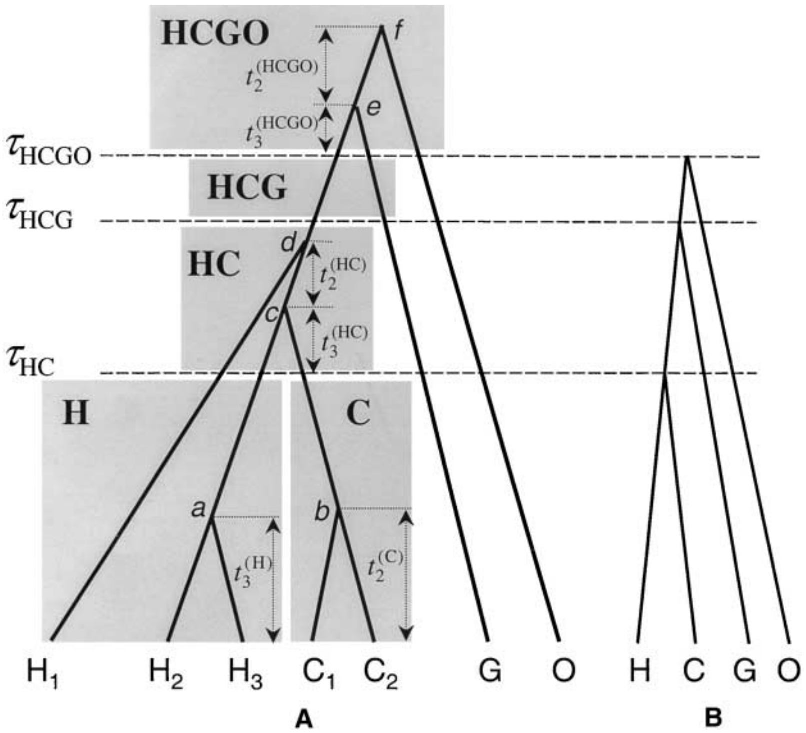

Figure 1.—(A) A gene tree for a locus with three humans (H), two chimpanzees (C), one gorilla (G), and one orangutan (O) for derivation of the probability distribution of gene trees and coalescent times. Both the speciation times (τ_{HC}, τ_{HCG}, and τ_{HCGO}) and the coalescent times (the *t*'s) are measured by the expected number of mutations per site. Coalescent processes in the five populations (denoted H, C, HC, HCG, and HCGO, shaded) have different population size parameters (θ_H, θ_C, θ_{HC}, θ_{HCG}, and θ_{HCGO}). (B) The tree topology of the species tree is assumed known.

$$f(D|G) = \prod_i f(D_i|G_i). \qquad (3)$$

Bayes inference is based on the joint conditional distribution

$$f(\Theta,\ G|D) \propto f(D|G)\ f(G|\Theta)\ f(\Theta). \qquad (4)$$

For example, the posterior density of Θ is given by

$$f(\Theta|D) = \int f(\Theta,\ G|D)\,dG, \qquad (5)$$

where the integration represents summation over all possible gene tree topologies and integration over the coalescent times at each locus.

We construct a Markov chain, whose states are (Θ, G) and whose stationary distribution is $f(\Theta,\ G|D)$. A Metropolis-Hastings algorithm (Metropolis *et al.* 1953; Hastings 1970) is used. Given the current state of the chain (Θ, G), a new state (Θ^*, G^*) is proposed through a proposal density, $q(\Theta^*,\ G^*|\Theta,\ G)$, and is accepted with probability

$$\begin{aligned} R &= \min\left\{1,\ \frac{f(\Theta^*,\ G^*|D)}{f(\Theta,\ G|D)} \times \frac{q(\Theta,\ G|\Theta^*,\ G^*)}{q(\Theta^*,\ G^*|\Theta,\ G)}\right\} \\ &= \min\left\{1,\ \frac{f(D|G^*)f(G^*|\Theta^*)f(\Theta^*)}{f(D|G)f(G|\Theta)f(\Theta)} \times \frac{q(\Theta,\ G|\Theta^*,\ G^*)}{q(\Theta^*,\ G^*|\Theta,\ G)}\right\}. \end{aligned} \qquad (6)$$

If the new state is accepted, the chain moves to the new state (Θ^*, G^*). Otherwise the chain stays in the old state (Θ, G). The challenge of the MCMC algorithm and the focus of this article is to derive the prior distribution, $f(G|\Theta)$, and to implement an efficient proposal algorithm and calculate the proposal ratio, $q(\Theta,\ G|\Theta^*,\ G^*)/q(\Theta^*,\ G^*|\Theta,\ G)$.

The proposal density q can be rather flexible as long as it specifies an aperiodic and irreducible Markov chain. The algorithm we implemented cycles through several steps, with each step updating some variables. Step 1 changes the age of an internal node in each gene tree without changing the gene tree topology or speciation times. Step 2 cycles through all loci and, at each locus, changes the gene tree topology by pruning a subtree and then regrafting it back onto a feasible branch. Step 3 updates the θ's. Step 4 cycles through all speciation times (τ's) in the species tree and modifies each. This step also uses a "rubber-band" algorithm to jointly modify the ages of nodes in each gene tree such that the coalescence times on the gene trees remain compatible with the modified species divergence times. Step 5 is a mixing step, in which all coalescent times in the gene trees and all species divergence times are multiplied by the same constant. The details of the algorithm are given in the appendix.

Distribution of the gene genealogy derived from censored coalescent processes: The prior probability, $f(G_i|\Theta)$, of any gene tree and its coalescent times at a locus are specified by the coalescent processes in the different populations in the species tree. The theory applies to any gene tree, but is best explained with an example, for which we use the gene tree of Figure 1. Five populations, H, C, HC, HCG, and HCGO, are considered. We use HC to represent the population ancestral to H, and C. Yang (2002; see also Takahata *et al.* 1995) derived the joint prior distribution $f(G_i|\Theta) =$

TABLE 1

Descriptions of the five populations (coalescent processes) represented in Figure 1, for deriving the probability of the gene tree and coalescent times

Population	θ	In (m)	Out (n)	Duration (τ)	Coalescent times
H	θ_{H}	3	2	τ_{HC}	$t_3^{(\mathrm{H})}$
C	θ_{C}	2	1	τ_{HC}	$t_2^{(\mathrm{C})}$
HC	θ_{HC}	3	1	$\tau_{\mathrm{HCG}} - \tau_{\mathrm{HC}}$	$t_3^{(\mathrm{HC})}$, $t_2^{(\mathrm{HC})}$
HCG	θ_{HCG}	2	2	$\tau_{\mathrm{HCGO}} - \tau_{\mathrm{HCG}}$	None
HCGO	θ_{HCGO}	3	1	∞	$t_3^{(\mathrm{HCGO})}$, $t_2^{(\mathrm{HCGO})}$

$f(T_i, \mathbf{t}_i|\Theta)$ for three species by considering the marginal probability of the tree topology T_i and the conditional distribution of the coalescent times given the topology; that is, $f(T_i, \mathbf{t}_i|\Theta) = f(T_i|\Theta)\ f(\mathbf{t}_i|T_i, \Theta)$. This strategy is not workable for larger species trees because of the increased number of tree topologies and the high dimension of the integral in deriving $f(T_i|\Theta)$. Here we derive the joint distribution $f(T_i, \mathbf{t}_i|\Theta)$ directly.

Note that two sequences from different species can coalesce only in populations that are ancestral to the two species. For example, sequences H_1 and G can coalesce in populations HCG or HCGO, but not in populations H or HC. The coalescent processes in different populations are independent. For each population, we trace the genealogy backward in time, until the end of the population at time τ, and record the number of lineages (m) entering the population and the number of lineages leaving it (n). For example, $m = 3$, $n = 2$, and $\tau = \tau_{\mathrm{HC}}$, for population H (Table 1). Such a coalescent process may be termed a *censored* coalescent process since the process is terminated before it is complete. When $n > 1$, the genealogical tree in the population consists of n disconnected subtrees or lineages.

Within each population, we measure time in units of $2N$ generations and further multiply time by the mutation rate. With this scaling, coalescent times are measured by the expected number of mutations per site, and any two lineages in the sample coalesce at the rate $\theta/2$ (Hudson 1990). The waiting time t_j until the next coalescent event, which reduces the number of lineages from j to $j - 1$, has the exponential density

$$f(t_j) = \frac{j(j-1)}{2} \times \frac{2}{\theta}\exp\left\{-\frac{j(j-1)}{2} \times \frac{2}{\theta}t_j\right\},$$
$$j = m, m-1, \ldots, n+1. \tag{7}$$

If $n > 1$, we have to consider the probability that no coalescent event occurred between the last coalescent event and the end of the population at time τ, that is, during the time interval $\tau - (t_m + t_{m-1} + \ldots + t_{n+1})$. This probability is $\exp\{-(n(n-1)/\theta)[\tau - (t_m + t_{m-1} + \ldots + t_{n+1})]\}$ and is 1 if $n = 1$. In addition, to derive the probability of a particular gene tree topology in the population, note that if a coalescent event occurs in a sample of j lineages, the probability that a particular pair of lineages coalesce is $1/\binom{j}{2} = 2/j(j-1)$, $j = m, m-1, \ldots, n+1$.

Multiplying those probabilities together, we obtain the joint probability distribution of the gene tree topology in the population and its coalescent times t_m, t_{m-1}, $\ldots$, t_{n+1} as

$$\prod_{j=n+1}^{m}\left[\frac{2}{\theta}\exp\left\{-\frac{j(j-1)}{\theta}t_j\right\}\right]$$
$$\times \exp\left\{-\frac{n(n-1)}{\theta}(\tau - (t_m + t_{m-1} + \ldots + t_{n+1}))\right\}. \tag{8}$$

The probability of the gene tree and coalescent times for the locus is the product of such probabilities across all the populations. Thus, for the gene genealogy of Figure 1, we have

$$f(G_i|\Theta) = [2/\theta_{\mathrm{H}}\exp\{-6t_3^{(\mathrm{H})}/\theta_{\mathrm{H}}\}\exp\{-2(\tau_{\mathrm{HC}} - t_3^{(\mathrm{H})})/\theta_{\mathrm{H}}\}]$$
$$\times [2/\theta_{\mathrm{C}}\exp\{-2t_2^{(\mathrm{C})}/\theta_{\mathrm{C}}\}]$$
$$\times [2/\theta_{\mathrm{HC}}\exp\{-6t_3^{(\mathrm{HC})}/\theta_{\mathrm{HC}}\} \times 2/\theta_{\mathrm{HC}}\exp\{-2t_2^{(\mathrm{HC})}/\theta_{\mathrm{HC}}\}]$$

$\times$ [exp{−2(~~$\tau_{\mathrm{HCG}} - \tau_{\mathrm{HC}}$~~ [redacted] $)/\theta_{\mathrm{HCG}}\}]$

$$\times [2/\theta_{\mathrm{HCGO}}\exp\{-6t_3^{(\mathrm{HCGO})}/\theta_{\mathrm{HCGO}}\} \times 2/\theta_{\mathrm{HCGO}}\exp\{-2t_2^{(\mathrm{HCGO})}/\theta_{\mathrm{HCGO}}\}]. \tag{9}$$

APPLICATION TO HOMINOID DATA

Data: We apply the new method to the following data, all composed of noncoding regions. Noncoding regions are preferable for this kind of analysis as they are likely to be evolving neutrally, less affected by background selection than, for example, silent sites in coding regions.

i. Chen and Li (2001) sequenced one individual from each of the four species, human, chimpanzee, gorilla, and orangutan, at 53 independent noncoding loci (contigs), with ~500 bp at each locus. Chen and Li's analysis using the tree-mismatch method estimated the population size for the common ancestor of humans and chimpanzees to be from 52,000 to 150,000. Maximum-likelihood (ML) analy-

TABLE 2

Prior and posterior distributions of parameters in the Bayes analysis of the 53 loci of Chen and Li (2001)

Parameter	(α, β)[a]	Prior	Posterior (53 loci)	Posterior (56 loci)
θ_{H}	(2, 2000)	0.001 (0.00012, 0.00279)		0.00057 (0.00039, 0.00076)
θ_{HC}	(2, 2000)	0.001 (0.00012, 0.00279)	0.00197 (0.00077, 0.00374)	0.00258 (0.00120, 0.00448)
θ_{HCG}	As above	As above	0.00342 (0.00219, 0.00487)	0.00387 (0.00262, 0.00536)
θ_{HCGO}	As above	As above	0.00198 (0.00023, 0.00446)	0.00235 (0.00033, 0.00467)
$\tau_{HCGO} - \tau_{HCG}$	(7.4, 1000)	0.0074 (0.00307, 0.01361)	0.00797 (0.00630, 0.00952)	0.00773 (0.00637, 0.00906)
$\tau_{HCG} - \tau_{HC}$	(4, 2500)	0.0016 (0.00044, 0.00351)	0.00118 (0.00047, 0.00211)	0.00134 (0.00063, 0.00213)
τ_{HC}	(20, 4000)	0.0050 (0.00305, 0.00742)	0.00481 (0.00406, 0.00554)	0.00432 (0.00369, 0.00495)

Both τ and θ are measured as the expected number of mutations per site.

[a] Parameters α and β are for the gamma prior, with the prior mean to be α/β.

[b] Mean and 2.5 and 97.5 percentiles of the prior or posterior distributions.

sis of the same loci using only the H-C-G sequences suggested smaller estimates of ~12,000–21,000 (Yang 2002). Here we use the data from all four species.

ii. Yu *et al.* (2001) sequenced ~10 kb at the region 1q24 from 61 humans, one chimpanzee, one gorilla, and one orangutan. This region was intended to be noncoding but was discovered to contain four exons (of 115, 155, 138, and 151 nucleotides long), which are removed before analysis.

iii. Makova *et al.* (2001) sequenced a region of ~6.6 kb at 16q24.3, located upstream from the melanocortin 1 receptor gene and containing its promoter, from 54 humans, one chimpanzee, one gorilla, and one orangutan. The orangutan sequence was incomplete and unavailable from GenBank. Only the human, chimpanzee, and gorilla sequences are used.

iv. Zhao *et al.* (2000) sequenced ~10 kb in the region 22q11.2 from 64 humans, one chimpanzee, and one orangutan. One human sequence (AF291608) appears to be corrupted, so only 63 human sequences are used.

In all data sets, most alignment gaps occur at the ends of the sequence and probably represent undetermined nucleotides. The three large loci (Zhao *et al.* 2000; Makova *et al.* 2001; Yu *et al.* 2001) involve many ambiguity nucleotides. These are included in the likelihood calculation (Yang 2000).

Two analyses are performed. The first estimates the ancestral population size and speciation time parameters θ_{HC}, θ_{HCG}, θ_{HCGO}, τ_{HCGO}, τ_{HCG}, and τ_{HC}, initially using the data set of Chen and Li (2001) at 53 loci and then including the data at the three other loci as well, in which case an additional parameter θ_{H} is also estimated. The results are presented in Table 2 under the headings "53 loci" and "56 loci," respectively. The second analysis uses only the human sequences at the three loci (Zhao *et al.* 2000; Makova *et al.* 2001; Yu *et al.* 2001) to estimate θ_{H} and t_{MRCA}, the time to the most recent common ancestor in the sample. The results are presented in Table 3.

Estimation of ancestral population sizes and speciation times: The gamma priors for the ancestral population size and speciation time parameters are specified on the basis of our expectations about those parameters (Table 2). For easy comparison, the same priors are used for parameters θ_{HC}, θ_{HCG}, $(\tau_{HCG} - \tau_{HC})$, and τ_{HC} as in Yang (2002), although parameters θ_{HCGO} and $(\tau_{HCGO} - \tau_{HCG})$ are new. The gamma parameter α is chosen to be >1, so that the distribution peaks at a positive value instead of 0. The prior for each θ has the mean 0.001 and 95% of the density is in the interval (0.00012, 0.00279). We assume a generation time of $g = 20$ years and a neutral mutation rate of 10^{-9} mutations/site/year. Thus the population sizes have a prior mean of 12,500 with the 95% interval (1500, 34,800). The mean speciation times in the prior are 5 million years (MY) before present, 6.6 MY, and 14 MY for the H-C, HC-G, and HCG-O divergences, respectively.

We use 10,000 iterations as the burn-in and then take 1,000,000 samples, sampling every two iterations. The results are presented in Tables 2 and 3. The posterior distribution for θ_{HC} from the 53-loci data (Chen and Li 2001) indicates a population size of 24,600 with the 95% credibility interval (CI) of (9600, 46,800) for the H-C ancestor. These are larger than the Bayes estimates obtained from the H-C-G sequences only, which were 13,100 with the 95% CI (1700, 32,100; Yang 2002; Table 3). The size for population HCG has posterior mean 42,700 with the 95% CI (27,000, 60,900), which are also larger than those from the H-C-G sequences only (Yang 2002). The H-C divergence time is calculated to be ~4.8 MY with the 95% CI (4.1, 5.5). Those estimates seem too young, as current opinion appears to favor a date as old as 7 MY (Brunet *et al.* 2002). The gap between the H-C and HC-G divergences is estimated to be ~1.2 MY with the 95% CI (0.47, 2.11), smaller than the estimates from the H-C-G sequences only (Yang 2002). We also analyzed the H-C sequences only from the Chen and Li data and obtained estimates of θ_{HC} and τ_{HC} that are almost identical to estimates from the H-C-G sequences (results not shown). In sum, inclusion of the

TABLE 3

Bayes estimates from the human data

		Bayes: mean (95% CI)	
Region and length (reference)	n	θ_H	μt_{MRCA}
1q24, ~10 kb (Yu *et al.* 2001)	61	0.00035 (0.00017, 0.00062)	0.00031 (0.00015, 0.00055)
16q24.3, ~6.6 kb (Makova *et al.* 2001)	54	0.00071 (0.00038, 0.00116)	0.00075 (0.00042, 0.00121)
22q11.2, ~10 kb (Zhao *et al.* 2000)	63	0.00065 (0.00038, 0.00103)	0.00051 (0.00030, 0.00079)
Combined		0.00056 (0.00040, 0.00076)	0.00037 (0.00020, 0.00062)
			0.00060 (0.00036, 0.00094)
			0.00045 (0.00028, 0.00070)

orangutan has led to more recent estimates of speciation times and to large estimates of ancestral population sizes. The reasons for the differences are unclear, but they are not due to exclusion of alignment gaps in the analysis of Yang (2002).

We note strong negative correlations in the posterior density between τ's and θ's, especially between τ and θ for the population representing the root of the species tree (Table 4). For example, the correlation between τ_{HCGO} and θ_{HCGO} is -0.75. The joint posterior density for θ_{HCGO} and ($\tau_{HCGO} - \tau_{HCG}$) is shown in Figure 2, after kernel density smoothing (Silverman 1986).

Including the three large loci of Yu *et al.* (2001), Makova *et al.* (2001), and Zhao *et al.* (2000) leads to even younger estimates of speciation times and larger estimates of ancestral population sizes (Table 2; column labeled 56 loci). The H-C divergence time, estimated at 4.3 MY with a 95% CI (3.7, 5.0), appears too recent. The strong correlations between parameters θ and τ in the posterior distribution suggest that estimation of θ's is affected by uncertainties in the τ's. To alleviate such effects, we used a highly informative prior for τ_{HC} with $\alpha = 120$, $\beta = 20{,}000$, corresponding to a prior mean of 6 MY for the H-C divergence with the 95% prior interval (5.0 MY, 7.1 MY). The 53-loci data then give the posterior mean 5.3 MY with the CI (4.7 MY, 5.9 MY) for the H-C divergence. The posterior mean and the 95% CI for θ_{HC} are 0.0015 (0.0006, 0.0029), which correspond to an HC population size of 19,000 with the CI (7600, 36,600), 0.0034 (0.0022, 0.0048) for θ_{HCG}, and 0.0019 (0.0002, 0.0044) for θ_{HCGO}. The posterior means and 95% CIs for speciation times are 0.0079 (0.0062, 0.0094) for $\tau_{HCGO} - \tau_{HCG}$ and 0.0010 (0.0004, 0.0017) for $\tau_{HCG} - \tau_{HC}$. Those estimates appear more reasonable.

Application of this informative prior for τ_{HC} to the 56-loci data had a similar effect of reducing θ_{HC}. The posterior mean and 95% CI for θ_{HC} are 0.00201 (0.00088, 0.00352), which correspond to a mean N_{HC} of 25,000 with the CI (11,000, 44,000). Estimates for τ_{HC} are 0.00481 (0.00430, 0.00535). Estimates for θ_H are 0.00055 (0.00039, 0.00075), identical to those of Table 2. Estimates of other parameters are 0.00379 (0.00258, 0.00529) for θ_{HCG} and 0.00240 (0.00042, 0.00453) for θ_{HCGO}. The posterior means and 95% CIs for speciation times are 0.00757 (0.00629, 0.00887) for $\tau_{HCGO} - \tau_{HCG}$ and 0.00109 (0.00048, 0.00177) for $\tau_{HCG} - \tau_{HC}$.

Estimation of human population size θ_H and the t_{MRCA}: The human sequences at the three large loci (Zhao *et al.* 2000; Makova *et al.* 2001; Yu *et al.* 2001) are analyzed separately and then combined. The only parameter to be estimated is θ_H, and the same gamma prior $G(2, 2000)$ is used as in Table 2, which corresponds to a prior mean of $N_H = 12{,}500$ with the 95% interval (1500, 34,800). The results are shown in Table 3.

For locus 1q24 (Yu *et al.* 2001), the Bayes analysis suggests a posterior mean of 0.00035 with the 95% CI (0.00017, 0.00062) for θ_H. With the generation time $g = 20$ years and mutation rate $\mu = 10^{-9}$ substitutions/site/year, those estimates correspond to an average long-term human population size of only $N_H = 4400$ with the 95% CI (2100, 7700). Yu *et al.* (2001) suggested a lower mutation rate for the locus at $\mu = 0.74 \times 10^{-9}$ substitutions/site/year. Use of this rate gives the posterior mean 5900 with the CI (2800, 10,500). The t_{MRCA} has the posterior mean 0.31 MY with the CI (0.15, 0.55) if the mutation rate is $\mu = 10^{-9}$ or 0.42 MY with the CI (0.20, 0.74) if the mutation rate is $\mu = 0.74 \times 10^{-9}$. Yu *et al.* (2001) estimated θ_H using Watterson's method based on the number of segregating sites (Watterson 1975), Tajima's method (Tajima 1983), and Fu and Li's BLUE method (Fu 1994), either with or without the singletons removed. The estimates varied considerably among methods and are all much larger than the Bayes estimates obtained here. The estimate suggested by the authors was $\theta = 6.7/8991 = 0.00074$, twice as large as the Bayes mean and outside the 95% CI. With $\mu = 0.74 \times 10^{-9}$ used, the population size was estimated to be $N_H = 12{,}600$ (Yu *et al.* 2001). Similarly Yu *et al.*'s analysis estimated t_{MRCA} of the human sample to be ~1.5 MY, more than three times older than the Bayes estimates.

For locus 16q24.3 (Makova *et al.* 2001), the Bayes analysis suggests a posterior mean of $N_H = 8800$ with the 95% CI (5500, 15,000) if $g = 20$ years and $\mu = 10^{-9}$.

TABLE 4

Correlation coefficients between parameters in the posterior distribution for the 53-loci data of Table 1

	θ_{HC}	θ_{HCG}	θ_{HCGO}	τ_{HCGO}	τ_{HCG}
θ_{HCG}	−0.21				
θ_{HCGO}	0.02	0.00			
τ_{HCGO}	*0.01*	*0.05*	*−0.75*		
τ_{HCG}	*0.24*	*−0.48*	*−0.01*	0.14	
τ_{HC}	*−0.55*	*0.01*	*−0.03*	0.12	0.40

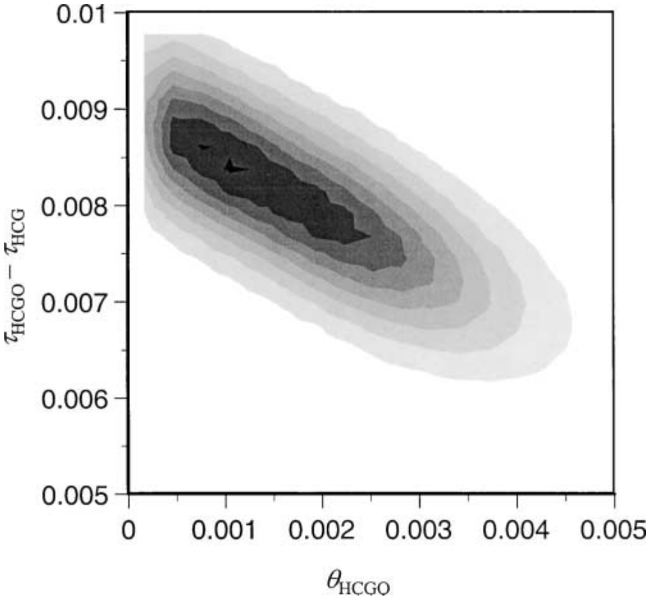

Figure 2.—Contour plot for the joint posterior density of θ_{HCGO} and $\tau_{HCGO} - \tau_{HCG}$. The correlation between the two parameters in the posterior is −0.71.

Makova *et al.* (2001) estimated a mutation rate of $\mu = 1.65 \times 10^{-9}$ for this locus. Use of this rate gives the posterior mean 5300 with the CI (3300, 9100) for N_H. The t_{MRCA} has the posterior mean 0.77 MY with the CI (0.44, 1.2) if the mutation rate is $\mu = 10^{-9}$ or 0.47 MY with the CI (0.27, 0.73) if the mutation rate is $\mu = 1.65 \times 10^{-9}$. Makova *et al.*'s (2001) estimates are $\theta = 13.7/6545 = 0.00209$, which is twice as large as the Bayes mean, and $N_H = 10{,}000$ for the human population size and $t_{MRCA} = \sim 1.5$ MY.

For locus 22q11.2 (Zhao *et al.* 2000), the Bayes analysis suggests a posterior mean of $N_H = 8100$ with the 95% CI (4700, 13,000), if $g = 20$ years and $\mu = 10^{-9}$. The t_{MRCA} has the posterior mean 0.51 MY with the CI (0.30, 0.79). Zhao *et al.*'s (2000) estimates of θ varied among methods and were all larger than the Bayes estimates. The population size N_H was estimated to be ~10,000–15,000, which is comparable with the Bayes estimate. However, t_{MRCA} was estimated to be ~1.3 MY, with a confidence interval of (0.71, 2.1), much larger than the Bayes estimates.

In sum, the Bayes estimates of θ_H and t_{MRCA} are considerably smaller than the estimates of Yu *et al.* (2001), Makova *et al.* (2001), and Zhao *et al.* (2000). As found by those authors, the estimation methods have a great impact. In all three data sets, there is an excess of rare mutants, such as singletons and doubletons (Zhao *et al.* 2000; Makova *et al.* 2001; Yu *et al.* 2001), which explains why estimates obtained using Watterson's method are often a few times larger than other estimates; this is the pattern expected for a recent population expansion. The exact reasons for the large differences are not entirely clear. One possible reason is the many ambiguity nucleotides in the human sequence data, which are properly dealt with in the Bayes and likelihood calculations but are typically removed in heuristic methods.

The three human loci are then combined in a Bayes analysis, with a single θ_H estimated (Table 3). The posterior mean $\theta_H = 0.00056$ with the 95% CI of (0.00040, 0.00076), corresponding to a population size of $N_H = 7000$ with the 95% CI of (5000, 9500), is an average across the three loci. However, the 95% CI is much narrower than those at individual loci, indicating the increased accuracy in the estimate in the combined analysis. The posterior means and CIs of the t_{MRCA} are similar to estimates from the separate analyses of the three loci (Table 3).

DISCUSSION

Validation of the MCMC algorithm and convergence monitoring: While MCMC provides a powerful framework for fitting sophisticated multiparameter models to heterogeneous data sets from multiple loci, MCMC algorithms are notoriously difficult to validate. MCMC implementations are notably more difficult to debug than maximum-likelihood programs. For example, in most numerical optimization algorithms for maximum-likelihood estimation, the likelihood always increases. However, an MCMC algorithm is stochastic and we cannot expect any summary statistics to increase or decrease monotonically. Second, a likelihood program converges to a fixed point (or points in the presence of multiple local optima). In contrast, convergence of an MCMC algorithm is to a distribution.

We used several strategies to validate the theory and implementation. For small data sets with only 2 or 3 species, quantities such as the probability of a particular gene tree topology and the expectations of coalescent times in the gene tree were calculated by both MCMC simulation and numerical integration using Mathematica. For larger species trees (with, say, 10 species), the MCMC algorithm was run without data [that is, by fixing $f(D|G) = 1$], and the resulting posterior distributions of parameters (Θ) were compared with the prior gamma

TABLE 5

Average size of the 95% CI

Parameter	Prior CI	(A) $L = 10$, $C = 10{,}000$	(B) $L = 20$, $C = 5000$	(C) $L = 100$, $C = 1000$	(D) $L = 500$, $C = 200$
θ_{HC}	0.0027	0.0016	0.0016	0.0014	0.0016
θ_{HCG}	0.0027	0.0017	0.0015	0.0013	0.0012
θ_{HCGO}	0.0027	0.0021	0.0023	0.0021	0.0019
$\tau_{HCGO} - \tau_{HCG}$	0.0105	0.0016	0.0016	0.0016	0.0015
$\tau_{HCG} - \tau_{HC}$	0.0031	0.0011	0.0010	0.0011	0.0012
τ_{HC}	0.0044	0.0008	0.0008	0.0009	0.0010

L is the number of loci and *C* is the sequence length. The number of replicates is (A) $r = 10$, (B) $r = 5$, (C) $r = 3$, and (D) $r = 2$.

distributions. In addition, a simulation program was written to generate testing data and to simulate the prior distribution of gene trees and coalescent times to validate the theory and the implementation.

For the purpose of convergence monitoring, we found it useful to run multiple chains and to monitor the values of parameters and the log likelihood over iterations (*e.g.*, Gelman and Rubin 1992). For the data analyzed in this study, our algorithm appeared to be fast to converge, even if poor starting points were used, but could be slow in mixing due to correlation between parameters (Table 4, Figure 2). A relatively short burn-in of 2000 or 5000 iterations appeared sufficient to bring all parameters to a reasonable range, with each iteration consisting of the five steps described in the appendix. After the burn-in, 10,000 samples, sampling every 2 iterations, produced stable estimates of posterior distributions. Results reported in Tables 2–4 were obtained from much longer runs.

Sampling strategies and accuracy of parameter estimation: A small simulation study was conducted to evaluate the information content in the data sets. Data are simulated using the species tree (((HC)G)O) with the parameter values $\theta_{HC} = 0.001$, $\theta_{HCG} = 0.001$, $\theta_{HCGO} = 0.001$, $\tau_{HCGO} = 0.014$, $\tau_{HCG} = 0.0066$, and $\tau_{HC} = 0.005$. The prior of Table 2 is used in the Bayes analysis. The fact that the prior means are equal to the true values of parameters suggests that the results are best-time results. Clearly the method will perform well if we have long sequences to reduce sampling errors in the gene tree and branch lengths (coalescent times) at each locus and also many loci to average over stochastic variations in the coalescent process among loci. However, given the total combined sequence length, it is not obvious whether it is better to have a few long sequences or many short ones. Thus we simulated a few cases in which the number of loci, *L*, and the number of nucleotides in the sequence at each locus, *C*, vary but the total sequence length from each species is fixed at $L \times C = 100{,}000$ (Table 5). The average posterior means of parameters (not shown) were found to be close to the true values of parameters in all the cases considered. We examined the size of the 95% credibility interval as an indication of the information content in the data. Surprisingly the results suggest very little difference among the different strategies (Table 5). Comparison of the posterior CIs with the prior intervals suggests that speciation times (τ's) are, in general, well estimated, with their CIs reduced by three to seven times. Parameters θ are estimated less well, with about twice the reduction in the CI for θ_{HC} and θ_{HCG}, while θ_{HCGO} is the most poorly estimated parameter.

Population size of the human-chimpanzee common ancestor: The human-chimpanzee ancestral population size, N_{HC} or θ_{HC}, has been of much recent interest (Takahata and Satta 2002; Wall 2003). As the size of modern humans has been consistently estimated to be ~10,000 (*e.g.*, Takahata *et al.* 1995; Ruvolo 1997; Zhao *et al.* 2000; Makova *et al.* 2001; Yu *et al.* 2001; Takahata and Satta 2002), which is surprisingly small given the widespread distributions of humans in the past 1–2 million years (Wall 2003), reliable estimation of the human-chimpanzee ancestral population size is essential to understand whether there has been a dramatic size reduction during human evolution (Hacia 2001; Kaessmann *et al.* 2001). Our likelihood and Bayes analyses of the data of Chen and Li (2001) produced estimates that are a few times smaller than estimates obtained from the same data using the tree-mismatch method (Yang 2002 and this study), demonstrating the importance of the estimation procedure.

However, the reliability of our estimates is affected both by the assumptions made in our model and by the quality of the data. We assumed that the evolutionary rate is the same both among sites within each locus and among different loci. Within-locus rate variation is not expected to be important because its effect is mainly on correction for multiple hits and because the sequences used are highly similar. This assumption can be relaxed, although at greater computational cost. Rate variation among loci should have a greater effect on estimation of ancestral population sizes (Yang 1997). It is straightforward to incorporate variable evolutionary rates across loci in the MCMC algorithm. However, the effect is less important when multiple-species data are analyzed simultaneously (Yang 2002). Another assumption we made is the absence of recombination within a locus. Takahata and Satta (2002) and Wall (2003) pointed out that recombination reduces variation among loci and leads to underestimates of N_{HC}. Furthermore, the data of Chen and Li (2001) may be somewhat atypical, since analyses of other genomic regions using the maximum-likelihood method of Takahata *et al.* (1995) have typically produced much larger estimates of N_{HC} (*e.g.*, Takahata *et al.* 1995; Takahata and Satta 2002). We note that the information content in the data analyzed here is low, as indicated by the sensitivity of our Bayes estimates to the prior, and that reliable estimates are possible only with accumulation of more genomic data and with implementation of more realistic models. We stress that our likelihood-based MCMC method makes an efficient use of the information in the data and provides a powerful framework for combining heterogeneous data sets from multiple loci.

Program availability: The program MCMCcoal, for Bayes MCMC analysis under coalescent models, is available at http://abacus.gene.ucl.ac.uk/software/MCMCcoal/. This can also be used to simulate sequence data sets.

This work was funded by a Biotechnology and Biological Sciences Research Council grant (31/G13580) to Z.Y. and grants from the National Institutes of Health (HG01988), the Canadian Institute of Health Research (MOP44064), the Peter Lougheed Foundation, and the Alberta Heritage Foundation for Medical Research to B.R.

LITERATURE CITED

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

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.

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

Brunet, M., F. Guy, D. Pilbeam, H. T. Mackaye, A. Likius *et al.*, 2002 A new hominid from the Upper Miocene of Chad, Central Africa. Nature **418:** 145–151.

Chen, F.-C., and W.-H. Li, 2001 Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Hum. Genet. **68:** 444–456.

Edwards, S. V., and P. Beerli, 2000 Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. Evolution **54:** 1839–1854.

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

Felsenstein, J., M. Kuhner, J. Yamato and P. Beerli, 1999 Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. IMS Lect. Notes Monogr. Ser. **33:** 163–185.

Fu, Y., 1994 Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. Genetics **138:** 1375–1386.

Gelman, A., and D. B. Rubin, 1992 Inference from iterative simulation using multiple sequences (with discussion). Stat. Sci. **7:** 457–511.

Griffiths, R. C., and S. Tavaré, 1994 Ancestral inference in population genetics. Stat. Sci. **9:** 307–319.

Hacia, J. G., 2001 Genome of the apes. Trends Genet. **17:** 637–645.

Hastings, W. K., 1970 Monte Carlo sampling methods using Markov chains and their application. Biometrika **57:** 97–109.

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, edited by D. J. Futuyma and J. D. Antonovics. Oxford University Press, New York.

Jukes, T. H., and C. R. Cantor, 1969 Evolution of protein molecules, pp. 21–123 in *Mammalian Protein Metabolism*, edited by H. N. Munro. Academic Press, New York.

Kaessmann, H., V. Wiebe, G. Weiss and S. Paabo, 2001 Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nat. Genet. **27:** 155–156.

Makova, K. D., M. Ramsay, T. Jenkins and W. H. Li, 2001 Human DNA sequence variation in a 6.6-kb region containing the melanocortin 1 receptor promoter. Genetics **158:** 1253–1268.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, 1953 Equations of state calculations by fast computing machines. J. Chem. Physiol. **21:** 1087–1092.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nielsen, R., and J. Wakeley, 2001 Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics **158:** 885–896.

Ruvolo, M., 1997 Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. Mol. Biol. Evol. **14:** 248–265.

Silverman, B. W., 1986 *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

Stephens, M., and P. Donnelly, 2000 Inference in molecular population genetics (with discussions). J. R. Stat. Soc. B **62:** 605–655.

Swofford, D. L., G. J. Olsen, P. J. Waddell and D. M. Hillis, 1996 Phylogeny inference, pp. 411–501 in *Molecular Systematics*, edited by D. M. Hillis, C. Moritz and B. K. Mable. Sinauer Associates, Sunderland, MA.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

Takahata, N., and Y. Satta, 2002 Pre-speciation coalescence and the effective size of ancestral populations, pp. 52–71 in *Developments in Theoretical Population Genetics*, edited by M. Slatkin and M. Veuille. Oxford University Press, Oxford.

Takahata, N., Y. Satta and J. Klein, 1995 Divergence time and population size in the lineage leading to modern humans. Theor. Popul. Biol. **48:** 198–221.

Wall, J. D., 2003 Estimating ancestral population sizes and divergence times. Genetics **163:** 395–404.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

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

Wilson, I. J., M. E. Weal and D. J. Balding, 2003 Inference from DNA data: population histories, evolutionary processes and forensic match probabilities. J. R. Stat. Soc. A **166:** 155–201.

Wu, C.-I, 1991 Inferences of species phylogeny in relation to segregation of ancient polymorphisms. Genetics **127:** 429–435.

Yang, Z., 1997 On the estimation of ancestral population sizes. Genet. Res. **69:** 111–116.

Yang, Z., 2000 Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. J. Mol. Evol. **51:** 423–432.

Yang, Z., 2002 Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics **162:** 1811–1823.

Yu, N., Z. Zhao, Y. X. Fu, N. Sambuughin, M. Ramsay *et al.*, 2001 Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. Mol. Biol. Evol. **18:** 214–222.

Zhao, Z., L. Jin, Y. X. Fu, M. Ramsay, T. Jenkins *et al.*, 2000 Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. Proc. Natl. Acad. Sci. USA **97:** 11354–11358.

Communicating editor: N. Takahata

APPENDIX: PROPOSAL STEPS IN THE MCMC

The MCMC algorithm implemented in this article involves several proposal steps, each of which updates some parameters in the Markov chain. The main problem is to combat the constraints posed by the speciation times while updating coalescent times in the gene tree and vice versa. The algorithm is tedious. The details follow.

Step 1. Updating the coalescent times at internal nodes in a gene tree: This step cycles through all loci and, for each locus, through all internal nodes in the gene tree to propose changes to the node ages (coalescent times). The age of only one internal node is changed at a time, and the gene tree topology remains intact. First the lower and upper bounds for the new age are determined by examining the current values of speciation dates and the ages of the mother and daughter nodes in the gene tree. A sliding window is then used to propose the new age; that is,

$$t_j^* = U(t_j - \varepsilon_1/2,\ t_j + \varepsilon_1/2), \tag{A1}$$

where $U(a, b)$ is a random variable from the uniform distribution in the interval (a, b), and ε_1 is the window size, which is adjustable. If the new age is outside the feasible range, the excess is reflected back into the interval. As there are always the same number of routes from t_j to t_j^* as from t_j^* to t_j, the proposal ratio is 1. The acceptance ratio is

$$R = \min\left\{1, \frac{f(D_i|G_i^*)f(G_i^*|\Theta)}{f(D_i|G_i)f(G_i|\Theta)}\right\}. \tag{A2}$$

Step 2. Subtree pruning and regrafting in the gene tree: This step cycles through nodes in each gene tree (except the root), removes the subtree represented by the node (which includes the node and all its descendent nodes), and then reattaches it to the remaining gene tree (Figure A1A). The step changes the gene tree topology and the coalescent time for the mother node. The feasible range for the new age of the mother node is determined on the basis of the current values of the speciation times and the age of concerned node in the gene tree. Then a random age is chosen by using a sliding window around the current age,

$$t_j^* = U(t_j - \varepsilon_2/2,\ t_j + \varepsilon_2/2), \tag{A3}$$

where the window size ε_2 can be fine tuned. If the new age is outside the feasible range, the excess is reflected back. Next the feasible branches in the gene tree at which the mother node can join are counted, and one of them is chosen at random. If m feasible lineages (to which the mother node can be attached) are in the current gene tree and n feasible lineages are in the proposed gene tree, the proposal ratio is n/m. Thus

$$R = \min\left\{1, \frac{f(D_i|G_i^*)f(G_i^*|\Theta)}{f(D_i|G_i)f(G_i|\Theta)} \times \frac{n}{m}\right\}. \tag{A4}$$

Note that this is the subtree-pruning and regrafting (SPR) algorithm used in phylogenetic tree search (Swofford *et al.* 1996), except that the gene tree is rooted, the age of the node is constrained, and the subtree can be attached to only some of the branches. An example is shown in Figure A1A, where the subtree (H_2, H_3) at node a is pruned and then reattached. The proposal also changes the age of the mother node c. Let the age of node a be t_a. The feasible range of t_c is (t_a, ∞). A new age t_c^* is proposed according to Equation A3, which happens to be in population HCG. There are $n = 2$ feasible branches (e–d and e–G) that the subtree can attach to at t_c^*, and one of them is chosen at random. In the current gene tree, the subtree can attach to $m = 2$ lineages (d–H_1 and d–b) at the current age t_c.

We also implemented a version of this proposal in which only the tips are pruned and regrafted. This was found to be effective for small gene trees, such as in the data of Chen and Li (2001), but is inefficient for large gene trees.

Step 3. Updating population size parameters: This step updates the population size parameters (θ's) one by one. A sliding window is used to propose a new value; that is, $\theta_j^* = U(\theta_j - \varepsilon_3/2,\ \theta_j + \varepsilon_3/2)$, where ε_3 is the window size. If $\theta_j^* < 0$, it is reset to $-\theta_j^*$. The proposal ratio is 1, and the acceptance ratio is

$$R = \min\left\{1, \frac{f(G|\Theta^*)f(\theta_j^*)}{f(G|\Theta)f(\theta_j)}\right\}. \tag{A5}$$

Step 4. Updating speciation times in the species tree: This step cycles through the speciation times, that is, ages at internal nodes of the species tree. The age τ at any internal node is bounded upward by the speciation time of its mother node and downward by the ages of its daughter nodes. Let the interval be (τ_L, τ_U). A sliding window is used to propose a new age,

$$\tau^* = U(\tau - \varepsilon_4/2,\ \tau + \varepsilon_4/2), \tag{A6}$$

where ε_4 is the adjustable window size. If the new age is outside the range (τ_L, τ_U), the excess is reflected back. To maintain the compatibility of the gene trees with the species tree, we change the ages of the affected nodes in the gene tree at each locus. A node in the gene tree is affected if its age is in the interval (τ_L, τ_U) and if it is in the population(s) represented by the concerned node in the species tree or its two daughter nodes.

Our calculation of the new ages for the affected nodes in the gene tree mimics the movements of marks (nodes in the gene tree) on a rubber band when its two ends are fixed (at τ_L and τ_U) and when the rubber is held at a fixed point (τ) and pulled slightly to one end (Figure A1B). The marks will move relative to the two ends when the rubber expands on one side of the holding point and shrinks on the other. If the node age in the gene tree $t > \tau$, the new age is given by $(\tau_U - t^*)/(\tau_U - t) = (\tau_U - \tau^*)/(\tau_U - \tau)$; that is,

$$t^* = \tau_U - \frac{(\tau_U - \tau^*)}{(\tau_U - \tau)}(\tau_U - t), \quad \text{for } t > \tau. \tag{A7}$$

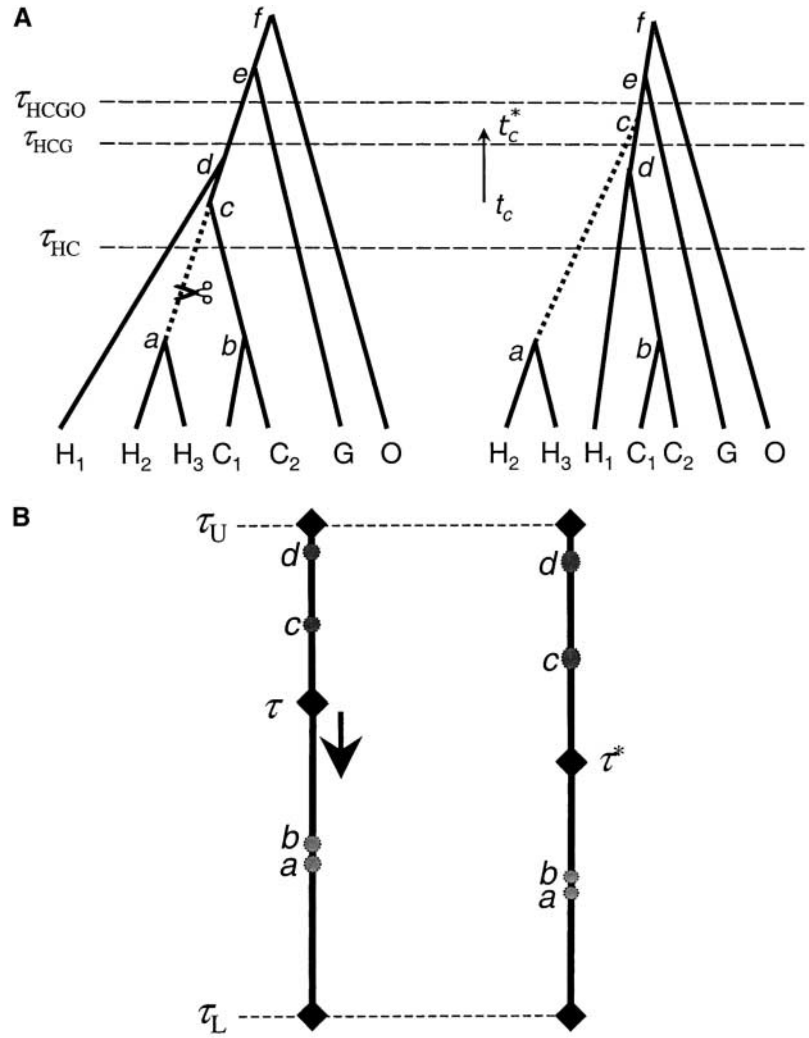

FIGURE A1.—(A) Subtree pruning and regrafting algorithm to update the gene tree topology. The subtree (H_2, H_3), represented by node a, is pruned by cutting branch a–c. The age t_c of the mother node is changed, and the subtree is regrafted to the gene tree at a feasible branch. (B) Rubber-band algorithm for updating a species divergence time τ, bounded by τ_L and τ_U. In the example, τ_{HC} is updated between $\tau_L = 0$ and $\tau_U = \tau_{HCG}$, and the four coalescent times correspond to nodes a, b, c, and d in the gene tree of Figure 1. When the proposal changes τ to τ^*, ages of nodes a, b, c, and d are also changed.

If the node age $t \le \tau$, the new age is given by $(t^* - \tau_L)/(t - \tau_L) = (\tau^* - \tau_L)/(\tau - \tau_L)$; that is,

$$t^* = \tau_L + \frac{(\tau^* - \tau_L)}{(\tau - \tau_L)}(t - \tau_L), \quad \text{for } t \le \tau. \qquad \text{(A8)}$$

If the node in the species tree is the root so that $\tau_U = \infty$, all node ages are changed relative to the lower bound (Equation A8).

An example is shown in Figure A1B, where τ_{HC} of Figure 1 is updated. The range is $\tau_L = 0$ and $\tau_U = \tau_{HCG}$. The ages of nodes a, b, c, and d in the gene tree should be changed as well to maintain compatibility between the gene tree and proposed species tree. Nodes a and b, which are younger than τ_{HC}, are repositioned relative to the lower bound τ_L according to Equation A8, while nodes c and d, which are older than τ_{HC}, are repositioned relative to the upper bound τ_U according to Equation A7.

Suppose m node ages are changed relative to the upper bound τ_U (using Equation A7) and n node ages are changed relative to the lower bound τ_L (using Equation A8) across all loci. To derive the proposal ratio, we apply the following transform: $y_0 = \tau$, $y_j = (\tau_U - t_j)/(\tau_U - \tau)$ for each of the m nodes with $t_j > \tau$, and $y_k = (t_k - \tau_L)/(\tau - \tau_L)$ for each of the n nodes with $t_k < \tau$. Note that only y_0 is changed while all other $m + n$ variables y_j and y_k remain the same in the proposal. The proposal ratio in the transformed variables is 1. The proposal ratio in the original variables is easily derived as $((\tau_U - \tau^*)/(\tau_U - \tau))^m((\tau^* - \tau_L)/(\tau - \tau_L))^n$. The acceptance ratio is thus

$$R = \min\left\{1, \frac{f(D|G^*)f(G^*|\Theta^*)}{f(D|G)f(G|\Theta)} \times \frac{f(\tau^*)}{f(\tau)} \times \left(\frac{\tau_U - \tau^*}{\tau_U - \tau}\right)^m \left(\frac{\tau^* - \tau_L}{\tau - \tau_L}\right)^n\right\}. \qquad \text{(A9)}$$

Step 5. Mixing step: A mixing step is found to be effective in speeding up convergence, especially from a poor starting point. The gene tree topologies remain unchanged, but all parameters in the model (θ's and τ's) and node ages (coalescent times) in each gene tree are multiplied by a constant

$$c = e^{\varepsilon_5(r-0.5)}, \qquad \text{(A10)}$$

where r is a random number from $U(0, 1)$ and $\varepsilon_5 > 0$ is a small fine-tuning parameter. The proposal ratio is

c^n, where n is the total number of variables updated. The acceptance ratio is

$$R = \min\left\{1, \frac{f(D|G^*)f(G^*|\Theta^*)}{f(D|G)f(G|\Theta)} \times \frac{f(\Theta^*)}{f(\Theta)} \times c^n\right\}. \tag{A11}$$

To overcome the strong correlation between parameters θ and τ (Table 4; see also Yang 2002), the mixing step is modified at an early stage of the MCMC run, say, when 10% of the samples have been taken, making use of the correlation coefficients between parameters calculated during the MCMC up to that point. For each parameter θ, its strongest correlation with parameters τ is found. If that correlation is <0.2, θ is not changed. Otherwise θ is multiplied by c if the correlation is positive or divided by c if the correlation is negative. Thus with the correlation coefficients of Table 4, all the θ parameters are divided by c. The proposal ratio for the modified algorithm is c^{m-n}, where m is the total number of parameters multiplied by c and n is the total number of parameters divided by c.

Derivation of the proposal ratio for the rubber band algorithm

Ziheng Yang
5 October 2012

The following is a detailed derivation of the proposal ratio (Hastings ratio) in equation (A9) on page 1655 of Rannala & Yang (2003 *Genetics* **164**:1645-1656) for the rubber band algorithm. The notation here is heuristic and is largely consistent with RY03. Let $\mathbf{t}_j$ be the set of m node ages on the gene tree that are older than the current τ; those are altered according to equation A7 in RY03. Let $\mathbf{t}_k$ be the set of n node ages on the gene tree that are younger than the current τ; those are altered according to equation A8. Note that the move changes τ, $\mathbf{t}_j$, and $\mathbf{t}_k$. We collect those into a vector $\mathbf{x} = \{\tau, \mathbf{t}_j, \mathbf{t}_k\}$, of size $(1 + m + n)$. Asterisks mean new proposed values. As stated in equation (A9), the proposal ratio is

$$\frac{q(\mathbf{x}\,|\,\mathbf{x}^*)}{q(\mathbf{x}^*\,|\,\mathbf{x})} = \frac{q(\tau,\mathbf{t}_j,\mathbf{t}_k\,|\,\tau^*,\mathbf{t}_j^*,\mathbf{t}_k^*)}{q(\tau^*,\mathbf{t}_j^*,\mathbf{t}_k^*\,|\,\tau,\mathbf{t}_j,\mathbf{t}_k)} = \left(\frac{\tau_U-\tau^*}{\tau_U-\tau}\right)^m \left(\frac{\tau^*-\tau_L}{\tau-\tau_L}\right)^n. \tag{1}$$

We show this below, by deriving $q(\mathbf{x}^*\,|\,\mathbf{x}) = q(\tau^*,\mathbf{t}_j^*,\mathbf{t}_k^*\,|\,\tau,\mathbf{t}_j,\mathbf{t}_k)$.

The move is one-dimensional, along a curve in the $(1 + m + n)$-dimensional space. As in RY03, define a transform $\mathbf{y}(\mathbf{x})$, with $\mathbf{y} = \{y_0, \mathbf{y}_j, \mathbf{y}_k\}$, as follows

$$\begin{aligned} y_0 &= \tau, \\ y_j &= \frac{\tau_U - t_j}{\tau_U - \tau}, \qquad j = 1, 2, \ldots, m, \\ y_k &= \frac{t_k - \tau_L}{\tau - \tau_L}, \qquad k = 1, 2, \ldots, n \end{aligned} \tag{2}$$

The Jacobi matrix of the transform is

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial(y_0,\mathbf{y}_j,\mathbf{y}_k)}{\partial(\tau,\mathbf{t}_j,\mathbf{t}_k)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ \vdots & \cdots & 0 & 0 & 0 \\ \frac{\tau_U - t_j}{(\tau_U-\tau)^2} & -\frac{1}{\tau_U-\tau} & \ddots & 0 & 0 \\ \vdots & \cdots & \cdots & \ddots & 0 \\ -\frac{t_k-\tau_L}{(\tau-\tau_L)^2} & \cdots & \cdots & \cdots & \frac{1}{\tau-\tau_L} \end{bmatrix}, \tag{3}$$

so the absolute value of the Jacobi determinant is $\left|\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right| = \left(\frac{1}{\tau_U-\tau}\right)^m \left(\frac{1}{\tau-\tau_L}\right)^n$. (My notation here is awkward as I am trying not to introduce new variables. The matrix is lower-diagonal, and on the diagonal, $-1/(\tau_U - \tau)$ occurs m times, and $1/(\tau - \tau_L)$ occurs n times.)

Note that from $\mathbf{y}$ to $\mathbf{y}$*, only y_0 changes, so that

$$q(\mathbf{y}^*\,|\,\mathbf{y}) = \tfrac{1}{\varepsilon_4} \qquad \text{(from equation A6 in RY03)} \tag{4}$$

Then

$$\begin{aligned} q(\mathbf{x}^*\,|\,\mathbf{x}) &= q_{\mathbf{x}^*|\mathbf{y}}(\mathbf{x}^*\,|\,\mathbf{y}) & &\text{Conditioning on } \mathbf{x} \text{ is the same as conditioning on } \mathbf{y} \\ &= q_{\mathbf{y}^*|\mathbf{y}}(\mathbf{y}^*(\mathbf{x}^*)\,|\,\mathbf{y}) \cdot \left|\tfrac{\partial \mathbf{y}^*}{\partial \mathbf{x}^*}\right| & &\text{Derive the density of } \mathbf{x}^* \text{ as a function of } \mathbf{y}^* \\ &= \tfrac{1}{\varepsilon_4} \cdot \left(\tfrac{1}{\tau_U-\tau^*}\right)^m \left(\tfrac{1}{\tau^*-\tau_L}\right)^n. \end{aligned} \tag{5}$$

Similarly $q(\mathbf{x}\,|\,\mathbf{x}^*) = \frac{1}{\varepsilon_4} \cdot \left(\frac{1}{\tau_U-\tau}\right)^m \left(\frac{1}{\tau-\tau_L}\right)^n$. Equation (1) thus follows.

The above also constitutes a proof of Theorem 2 on page 313 of Yang (2006 *Computational Molecular Evolution*, OUP). Note that there $J(\mathbf{y})$ = $\left|\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right| = 1\Big/\left|\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right|$.