constructive comments and Olaf Bininda-Emonds and Mike Sanderson for providing a preprint of their simulation work.

REFERENCES

BARRETT, M., M. J. DONOGHUE, AND E. SOBER. 1991. Against consensus. Syst. Zool. 40:486–493.

BAUM, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon 41:3–10.

BAUM, B. R., AND M. RAGAN. 1993. Reply to A.G. Rodrigo's "A comment on Baum's method for combining phylogenetic trees". Taxon 42:637–640.

BININDA-EMONDS, O. R. P., AND H. N. BRYANT. 1998. Properties of matrix representation with parsimony analysis. Syst. Biol. 47:497–508.

BININDA-EMONDS, O. R. P., J. L. GITTLEMAN, AND A. PURVIS. 1999. Building large trees by combining phylogenetic information: A complete phylogeny of the extant Carnivora (Mammalia). Biol. Rev. 74:143–175.

BININDA-EMONDS, O. R. P., AND M. SANDERSON. 2001. Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. Syst. Biol. 50:565–579.

GORDON, A. D. 1986. Consensus supertrees: The synthesis of rooted trees containing overlapping sets of labelled leaves. J. Classif. 9:335–348.

KLUGE, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). Syst. Zool. 38:7–25.

LAPOINTE, F. J., AND G. CUCUMEL. 1997. The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. Syst. Biol. 46:306–312.

LIU, R. F., M. M. MIYAMOTO, N. P. FREIRE, P. Q. ONG, M. R. TENNANT, T. S. YOUNG, AND K. F. GUGEL. 2001. Molecular and morphological supertrees for eutherian (placental) mammals. Science 291:1786–1789.

PURVIS, A. 1995a. A composite estimate of primate phylogeny. Philos. Trans. R. Soc. London B 348:405–421.

PURVIS, A. 1995b. A modification to Baum and Ragan's method for combining phylogenetic trees. Syst. Biol. 44:251–255.

RAGAN, M. A. 1992. Phylogenetic inference based on matrix representation of trees. Mol. Phylogenet. Evol. 1:53–58.

SEMPLE, C., AND M. STEEL. 2000. A supertree method for rooted trees. Discrete Appl. Math. 105:147–158.

STRIMMER, K., AND A. VON HAESELER. 1996. Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. Mol. Biol. Evol. 13:964–969.

WILKINSON, M., J. L. THORLEY, D. T. J. LITTLEWOOD, AND R. A. BRAY. 2001. Towards a phylogenetic supertree for the Platyhelminthes? Pages 292–301 *in* Interrelationships of the Platyhelminthes (D. T. J. Littlewood and R. A. Bray, eds.). Chapman-Hall, London.

Received 24 April 2001; accepted 2 September 2001
Associate Editor: M. Sanderson

Syst. Biol. 51(1):155–165, 2002

Geographic Origin of Human Mitochondrial DNA: Accommodating Phylogenetic Uncertainty and Model Comparison

JOHN P. HUELSENBECK AND NIKITA S. IMENNOV

Department of Biology, University of Rochester, Rochester, New York 14627, USA; E-mail: johnh@brahms.biology.rochester.edu

Abstract.—Many biogeographic problems are tested on phylogenetic trees. Typically, the uncertainty in the phylogeny is not accommodated when investigating the biogeography of the organisms. Here we present a method that accommodates uncertainty in the phylogenetic trees. Moreover, we describe a simple method for examining the support for competing biogeographic scenarios. We illustrate the method using mitochondrial DNA sequences sampled from modern humans. The geographic origin of modern human mtDNA is inferred to be in Africa, although support for this hypothesis was ambiguous for data from an early paper.

Debate on the origin of anatomically modern humans has concentrated on two competing hypotheses. The "out-of-Africa hypothesis" argues that modern humans originated in Africa and then migrated to other parts of the world, replacing other species of *Homo* as they spread (Stringer and McKie, 1996). The "regional continuity hypothesis," on the other hand, argues for a single species of *Homo* widely spread throughout the Old World, with populations connected through gene flow (Wolpoff and Caspari, 1997). Anatomically modern humans are then thought to have originated over a wide geographic area and after any migration event from Africa. The usual approach taken in a molecular test of these competing hypotheses is to collect DNA

sequences, usually from a single nonrecombining locus, and then to reconstruct the phylogenetic history of the gene. If the most recent common ancestor (MRCA) of modern humans is reconstructed as coming from Africa, then the data are considered to be consistent with the out-of-Africa hypothesis. This approach has been applied numerous times, first on mitochondrial DNA (mtDNA) sampled from individuals from diverse geographic origins (Cann et al., 1987; Vigilant et al., 1991) and later with genetic data sampled from the Y chromosome (Underhill et al., 2000). With a few exceptions (e.g., see Adcock et al., 2001), the results appear to favor the out-of-Africa hypothesis.

Any test of the out-of-Africa hypothesis presents several difficult statistical challenges, reviewed in Penny et al. (1995). First, the datasets generated tend to be very large. The early paper by Vigilant et al. (1991), for example, included >130 sequences, which at the time represented a very large phylogenetic analysis. More recent datasets include >1,000 DNA sequences (Krings et al., 1997). Adequately exploring tree space by using a reasonable method of phylogenetic analysis remains very difficult for large datasets. Second, the uncertainty in the phylogenetic trees must be satisfactorily accommodated. Many of the phylogenetic analyses of DNA sequences sampled from modern humans have a large degree of uncertainty (Maddison et al., 1992); it is important, then, that a test of the out-of-Africa hypothesis not depend on any single tree being correct. Finally, it is not clear how to evaluate the support that an alignment of DNA sequences has for the out-of-Africa hypothesis. This paper is intended to describe one framework for investigating the out-of-Africa hypothesis, as well as related questions. The approach can be applied to large datasets, provides a framework for evaluating competing hypotheses, and addresses the shortcomings of earlier approaches while accommodating phylogenetic uncertainty. We apply the method to the original mtDNA sequences analyzed by Vigilant et al. (1991), plus the few additional sequences examined by Maddison et al. (1992). We show that, although the original Vigilant et al. (1991) study was consistent with an African origin for modern human mtDNA, the evidence for this hypothesis is not overwhelming. We also demonstrate how previous work can serve as the prior for further data analysis, allowing diverse types of information to be combined in a single study. We do this using a new dataset of 200 human sequences; a recent Neandertal sequence serves as the outgroup.

Accommodating Phylogenetic Uncertainty When Testing the Out-of-Africa Hypothesis

Testing the out-of-Africa hypothesis involves comparing two hypotheses: that modern humans originated in Africa, and that modern humans did not originate in Africa. These hypotheses are difficult to test directly with molecular data. What usually can be inferred is the geographic origin of the particular sequence that was examined, but not the geographic origin of a population. A sample of individuals sequenced for a nonrecombining genetic locus will yield a most recent common ancestor; the geographic origin of the sample can be inferred by various techniques, such as the parsimony method. Distinguishing between the out-of-Africa and the regional continuity hypotheses on the basis of genetic data from a single locus may be difficult. The regional continuity hypothesis, for example, is consistent with some portions of the human genome having common ancestor in Africa. However, as more loci are sampled, which hypothesis best explains the data should become more clear. In this paper, we concentrate on how to analyze the genetic data from a single locus. Moreover, we are interested in comparing the following hypotheses—M_1, M_2, M_3, M_4, and M_5,—that the common ancestor of the sample lived in Africa, Europe, the Americas, Asia, or Australia/Oceania, respectively.

We assume that an alignment of s DNA sequences, each c sites long, is provided. We also assume that the sequences do not recombine, as would be the case for genetic data sampled from mtDNA or from the Y chromosome (but see Awadalla et al., 1999). By assuming that recombination is not a factor, we can allow the same phylogenetic tree to apply to all of the sites in the sequence. Moreover, when the ancestral geographic region of the sequence is inferred, it will apply to the entire sequence. The aligned DNA sequences will be denoted $\boldsymbol{X} = \{x_{ij}\}$, where $i = 1, \ldots, s$ and $j = 1, \ldots, c$. We wish

to evaluate the posterior probability of each hypothesis conditional on the observed DNA sequences, $f(M_i \mid \boldsymbol{X})$. In this study, we calculate the posterior probability of a hypothesis by summing the posterior probabilities of all phylogenetic trees that are consistent with the hypothesis (Huelsenbeck et al., 2000). We use the parsimony method to reconstruct the ancestral geographic region for a tree. Each phylogenetic tree, τ, is either inconsistent with, consistent with, or ambiguously supports, an origin on continent i. Hence, the posterior probability of an origin on continent i is

$$f(M_i \mid \boldsymbol{X}) = \sum_{j=1}^{B(s)} I(\tau_j) f(\tau_j \mid \boldsymbol{X})$$

where $B(s)$ is the number of possible phylogenies for s sequences and $I(\tau_j)$ is an index variable that takes the values

$$I(\tau_j) = \begin{cases} 0, & \text{if the MRCA is not assigned to continent } i \\ 1/k, & \text{if the MRCA is ambiguously assigned to continent } i \\ 1, & \text{if the MRCA is assigned to continent } i \end{cases}$$

(k is the number of ambiguous ancestral areas). Here, $f(\tau_i \mid \boldsymbol{X})$ is the posterior probability of the ith phylogenetic tree. The posterior probability of a phylogeny is calculated using Bayes's rule

$$f(\tau_i \mid \boldsymbol{X}) = \frac{f(\boldsymbol{X} \mid \tau_i) f(\tau_i)}{\sum_{j=1}^{B(s)} f(\boldsymbol{X} \mid \tau_j) f(\tau_j)}$$

where $f(\boldsymbol{X} \mid \tau_i)$ is the likelihood of the ith phylogeny and $f(\tau_i)$ is the prior probability of the ith phylogeny (Li, 1996; Mau, 1996; Rannala and Yang, 1996; Mau and Newton, 1997; Yang and Rannala, 1997; Larget and Simon, 1999; Mau et al., 1999; Newton et al., 1999). In this paper, we consider all $B(s)$ trees to be equally probable a priori [i.e., $f(\tau_i) = \frac{1}{B(s)}$]. Knowledge of only the phylogeny of the species, however, is insufficient to allow calculation of the likelihood (Felsenstein, 1981). In addition to a phylogenetic tree, one must assume a stochastic model of DNA substitution to calculate the likelihood. A phylogenetic model incorporates information on the lengths of the branches on the ith phylogeny (ν_i) as well as information on the pattern of nucleotide substitution (parameters contained in a vector $\boldsymbol{\theta}$). The likelihood of the ith tree, then, is

$$f(\boldsymbol{X} \mid \tau_i) = \int_{\nu, \boldsymbol{\theta}} f(\boldsymbol{X} \mid \tau_i, \boldsymbol{\nu}_i, \boldsymbol{\theta}) f(\boldsymbol{\nu}_i, \boldsymbol{\theta})\, d\boldsymbol{\nu}_i\, d\boldsymbol{\theta}$$

where integration is over the space of the branch lengths and substitution model parameters. Likelihoods can be calculated under any of the standard models of DNA substitution, reviewed in Swofford et al. (1996). In this study, we use uninformative priors for the branch lengths and substitution model parameters over the range of reasonable values for these parameters. Specifically, we assume a uniform (0,10) prior on branch lengths, a uniform (0,20) prior on the rates of substitution, a uniform (0,50) prior on the shape parameter of the gamma distribution (for among-site rate variation), and a flat Dirichlet (1,1,1,1) prior on base frequencies. These priors were chosen because they are flat over biologically plausible values for the parameters. Moreover, we assume a general model of DNA substitution that allows each substitution type to have its own rate and allows base frequencies potentially to differ (i.e., the GTR model of DNA substitution; Tavaré, 1986). The rate of substitution from nucleotide i to nucleotide j is designated r_{ij}. Rate variation across sites is accommodated by assuming that the rate at a site is a random variable drawn from a gamma distribution with shape and scale parameters set to α (Yang, 1993, 1994). Specifically, we use the discrete approximation, suggested by Yang (1994), with four rate categories. The parameters of the substitution model, then, are $\boldsymbol{\theta} = \{r_{AC}, r_{AG}, r_{AT}, r_{CG}, r_{CT}, \pi_A, \pi_C, \pi_G, \alpha\}$ ($r_{ij} = r_{ji}$ and the substitution rates are relative to the rate between $G \leftrightarrow T$, $r_{GT} = 1$; $\pi_T = 1 - \pi_A - \pi_C - \pi_G$). A general model of DNA substitution was chosen because it accommodated several possible rate models. Inspection of posterior distributions for rate parameters, base frequencies, and among-site rate variation suggests that the general model was appropriate (Table 1).

One potential weakness of the approach we have described is its reliance on the parsimony method to reconstruct the geographic region of the MRCA of the sequences. The parsimony method can become unreliable when character transformations are frequent relative to branching events. Hence, another

TABLE 1. Parameter estimates of the substitution model. The columns indicate the parameter, mean, and 95% credible interval for the parameter. The parameters are V, the tree length; r_{ij}, rate of substitution between nucleotides i and j measured relative to the rate between G and T ($r_{GT} = 1$); π_i, base frequencies; and α, gamma shape parameter for among-site rate variation.

Parameter	Mean	Credible interval
V	1.81	(1.51, 2.30)
r_{AC}	2.53	(0.75, 5.55)
r_{AG}	31.54	(9.69, 64.68)
r_{AT}	2.19	(0.54, 5.12)
r_{CG}	2.48	(0.55, 6.04)
r_{CT}	48.71	(14.48, 98.01)
r_{GT}	1.0	
π_A	0.302	(0.277, 0.327)
π_C	0.347	(0.323, 0.372)
π_G	0.134	(0.117, 0.152)
π_T	0.217	(0.197, 0.237)
α	0.246	(0.211, 0.285)

assumption of this analysis is that migration events are rare compared with coalescence events. This assumption can be relaxed by assuming that the characters (continental area) change on the tree according to a stochastic process, such as the Markov–Bernoulli process; we do not do so in this analysis, however, because we do not expect character transformations to be frequent relative to branching events. In this case, character transformations represent migration events among continents. Continental migration is probably rare relative to coalescence events.

The posterior probability of a phylogenetic tree is almost impossible to evaluate analytically because it involves a summation over all possible trees and, for each tree, integration over the branch lengths and substitution model parameters. We use the program MrBayes 1.1 (Huelsenbeck and Ronquist, 2001) to approximate the posterior probabilities of trees. This program uses a variant of Markov chain Monte Carlo (MCMC) that is less prone to entrapment in local optima than is normal MCMC (Metropolis et al., 1953; Hastings, 1970; Geyer, 1991; Green, 1995). MCMC approximates the posterior probability of a phylogenetic tree by constructing a Markov chain that has as its state space the parameters of the phylogenetic model. The states of the chain are sampled as it runs. The proportion of the time that any phylogenetic tree is represented in the sample is a valid approximation of its posterior probability (see Tierney, 1994). The variant of MCMC we use, Metropolis-coupled Markov chain Monte Carlo (Geyer, 1991), runs several chains simultaneously. All but one of the chains are heated, meaning that they more easily explore the space of phylogenetic trees. A heated chain has steady-state distribution of $f(\tau \mid \boldsymbol{X})^{\beta}$, where β is a heating parameter. The heated chains better explore the space of trees because the tree landscape is flattened relative to the cold (or unheated; $\beta = 1$) chain. After all of the chains have been advanced one step, a swap of the states for two randomly chosen chains is attempted. This strategy allows the cold chain to jump a deep valley in the landscape of trees when a successful swap between the cold and a heated chain is made.

MCMC is not a maximization (or minimization) algorithm that may be familiar to the reader. Instead, MCMC is used to sample trees in proportion to their probabilities. Trees that are more probable will be represented more often in the sample of trees than will those that are poor descriptions of the data. The most probable tree, known as the maximum posterior probability (MAP) estimate, may not even find its way into the sample if the probabilities of the best trees are very similar. Previously, Penny et al. (1995) used a heuristic search method to explore tree space for the data from Vigilant et al. (1991). Specifically, they used the Great Deluge algorithm to explore the landscape of trees developed using the parsimony criterion. The advantage of the Penny et al. (1995) method, shared by our own, is that inferences are not based on any single tree. However, the Penny et al. (1995) approach has two problems. First, the optimality criterion, parsimony, is known to fail under certain branch length conditions, and the criterion's statistical basis is poor. Second, the Great Deluge strategy is not guaranteed to meet any particular sampling strategy for trees. The MCMC approach, on the other hand, can sample trees according to their posterior probabilities.

We analyzed the dataset of $s = 140$ mitochondrial sequences of $c = 1{,}139$ sites used in the study by Maddison et al. (1992). This study included 135 modern human sequences, 78 of which were from individuals with origins in Africa. The dataset also includes five chimpanzee sequences, which were used to root the trees. We ran two MCMC analyses for 2,000,000 generations

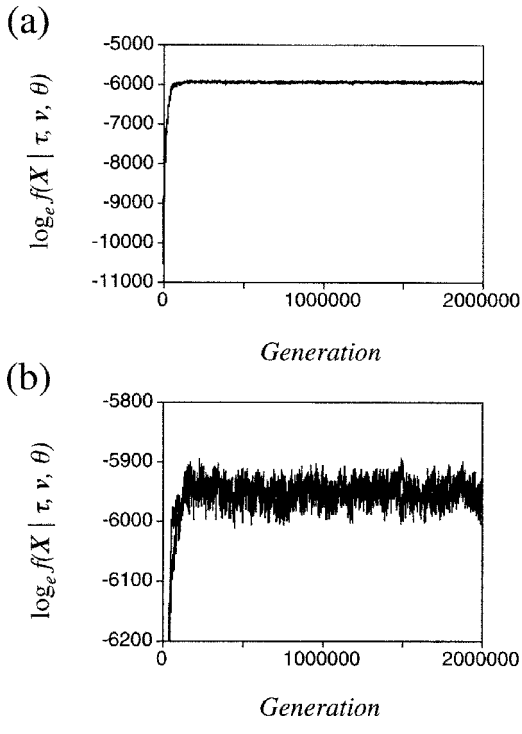

FIGURE 1. The log probability of the observed DNA sequences, ***X***, through time for both of the chains run in this study. Each chain started from different random trees and reached apparent stationarity by 100,000 generations. The samples taken from the first 500,000 generations were discarded as the burn-in for the chain, and inferences are based on samples from the remaining parts of the chain.

each. Each chain consisted of one cold and three heated chains and the Markov chains were started from independent random trees. Figure 1 shows the log probability of the data through time. Note that the chains initially started with trees that poorly explained the data but quickly found more reasonable trees. More importantly, both chains plateaued to the same log probability, suggesting that the chains have converged. This conclusion is supported by the fact that the inferences that would be drawn from the two chains are very similar. Figure 2, showing the posterior probability of individual clades found in both chains, illustrates that the posterior probabilities for the same clades are highly correlated.

The Markov chains were sampled every 100 generations, resulting in a total of 20,000 sampled trees from each chain. The first 5,000 trees were discarded from each as the "burn-in" (the portion of the chain that was sampled before stationarity was reached).

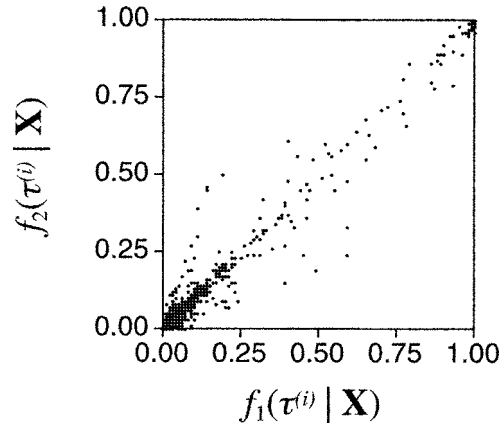

FIGURE 2. Correlation between the posterior probabilities of individual clades, $f(\tau^{(i)} \mid X)$, obtained from the separate Markov chains.

Inferences, then, were based on the 30,000 trees sampled from both chains. The MCMC strategy, when successfully completed, ensures that the trees were sampled in proportion to their posterior probability. Figure 3 shows the 50% majority rule consensus tree for the data of Maddison et al. (1992). The numbers at the interior branches of the tree represent the posterior probability that the clade is correct. Note that large portions of the tree are poorly supported, with posterior probabilities of <50% (the regions of the tree that contain large polytomies). Table 1 provides the estimates of the substitution parameters. The low support for many of the branches suggests that inferring the ancestral continental region for the sequences may be difficult. However, such is not the case. Although many different trees were sampled in the MCMC approach, some proportion of these trees will be consistent with a MRCA from Africa. We counted the number of the trees that were consistent with an African origin of the mtDNA by using the program PAUP* (Swofford, 1998). We rooted all of the trees, using the chimpanzees as the outgroup, and examined the parsimony reconstruction for the node that represented the MRCA of the modern human sequences. All equally parsimonious reconstructions at the ancestor of modern humans were calculated by using the Fitch (1971) algorithm. Of the 30,000 sampled trees, 27,601 were consistent with an African origin (i.e., the MRCA of the mtDNA was unambiguously reconstructed as residing in Africa), 1,154 were consistent with the alternative hypothesis, and

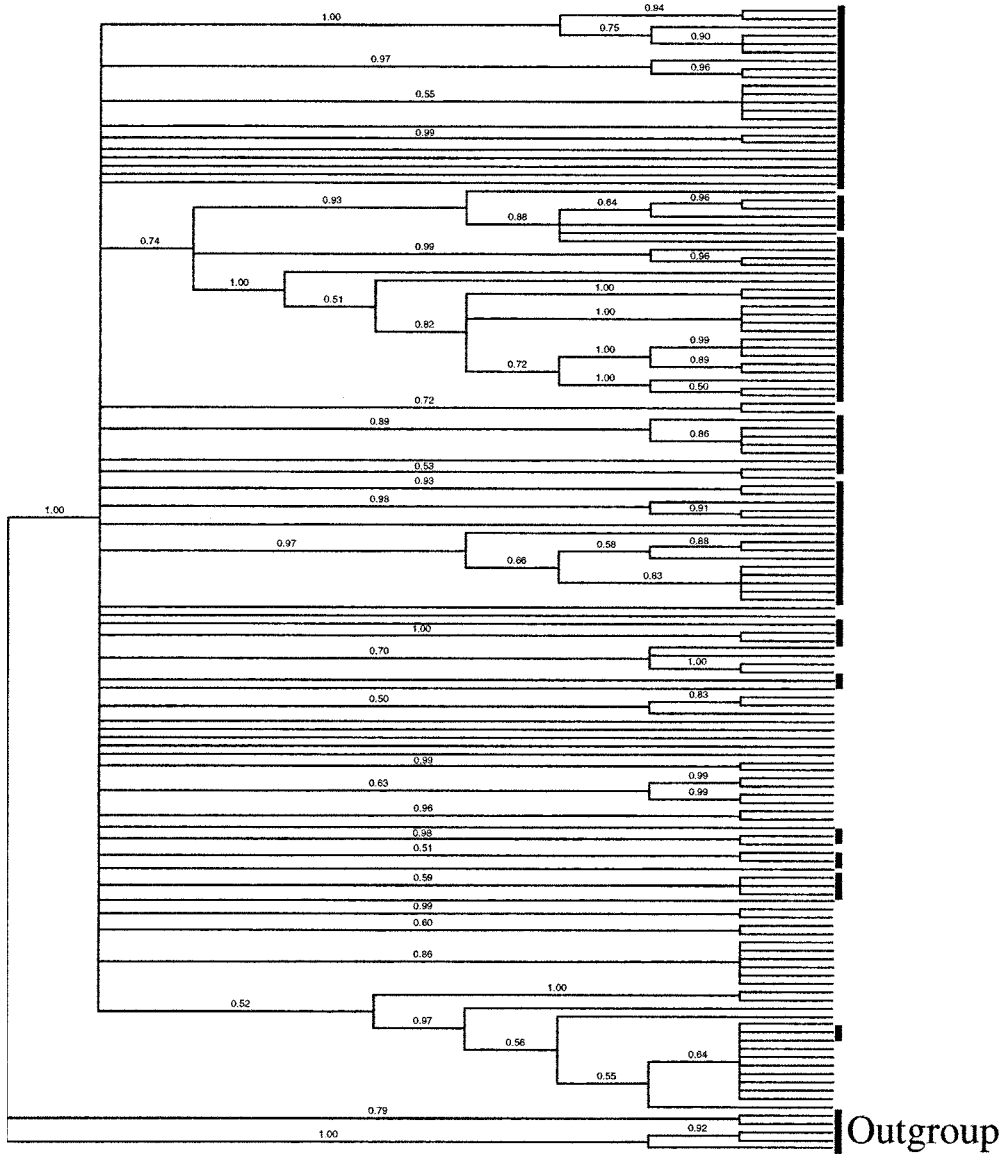

FIGURE 3. The 50% majority rule consensus tree of the 30,000 trees sampled in the MCMC analyses. Individuals from Africa are indicated by the vertical bars.

1,245 trees were ambiguous (i.e., reconstructions with two or more geographic areas were equally parsimonious). When the state assigned to the MRCA was ambiguous, we assigned equal weight to each of the possible reconstructions, which yielded the following posterior probabilities of the five hypotheses:

$$f(M_1 \mid \boldsymbol{X}) = 0.94$$
$$f(M_2 \mid \boldsymbol{X}) = 0.01$$
$$f(M_3 \mid \boldsymbol{X}) = 0.00$$
$$f(M_4 \mid \boldsymbol{X}) = 0.03$$
$$f(M_5 \mid \boldsymbol{X}) = 0.02$$

This appears to be strong support for an African origin for human mtDNA because $f(M_1 \mid \boldsymbol{X})$ is so much larger than the other possible root positions of the trees.

Model Comparison

What the above calculation fails to take into consideration is the prior probabilities of the five hypotheses. The posterior probability of the African origin hypothesis considers the probability of the hypothesis after the mtDNA sequences were observed. However, before consideration of the data, the decision to sample 78 African sequences and 57 non-African sequences induced a prior probability for an African origin of the mitochondrial sequences. We calculated the prior probability of the ith hypotheses as

$$f(M_i) = \sum_{j=1}^{B(s)} I(\tau_j) f(\tau_j)$$

Each of the $B(s)$ trees is considered to be a priori equally probable. Hence, we were able to approximate the prior probability of each hypothesis by generating a sample of 10,000 random trees. We evaluated whether each of these random trees was consistent with an origin for the MRCA on continent i. The prior probabilities of the hypotheses were found to be as follows:

$$f(M_1) = 0.80$$
$$f(M_2) = 0.05$$
$$f(M_3) = 0.00$$
$$f(M_4) = 0.08$$
$$f(M_5) = 0.07$$

Hence, even before analysis of the mtDNA sequence data, an African origin for the sequences was nearly four times as likely as a hypothesis that has human mtDNA originating elsewhere. The mitochondrial data changed our opinion about the merit of an African origin hypothesis. Before observing the data, the probability of this hypothesis is 0.80, whereas after the analysis, the probability is 0.94. In other words, the mtDNA sequences cause us to change our opinion to be more in favor of an African origin for modern human mtDNA.

The amount by which our opinion is changed can be measured as the change in the odds of the hypotheses. The posterior odds of the African origin hypothesis and its converse is

$$\underbrace{\frac{f(M_1 \mid \boldsymbol{X})}{1 - f(M_1 \mid \boldsymbol{X})}}_{\text{Posterior odds}} = \underbrace{\frac{f(\boldsymbol{X} \mid M_1)}{f(\boldsymbol{X} \mid M_I^C)}}_{\text{Bayes factor}} \times \underbrace{\frac{f(M_1)}{1 - f(M_1)}}_{\text{Prior odds}}$$

The Bayes factor, then, is the ratio of the posterior odds to the prior odds and measures "the change in the odds in favor of the hypothesis when going from the prior to the posterior" (Lavine and Schervisch, 1999:120). For the human mtDNA data, then,

$$\frac{f(\boldsymbol{X} \mid M_1)}{f(\boldsymbol{X} \mid M_I^C)} = \frac{0.94/0.06}{0.80/0.20} = 3.9$$

which means that the Vigilant et al. (1991) data should cause us to change our opinion by a factor of 4 in favor of an African origin for the mtDNA. The original Vigilant et al. (1991) data are consistent with the out-of-Africa hypothesis.

However, a Bayes factor (or likelihood ratio) of 3 or 4 is not typically considered overwhelming support for one hypothesis over another. Jeffreys (1935, 1961) argued that competing models be compared by using the Bayes factor and provided a table for interpreting Bayes factors, which was modified by Raftery (1995) as follows:

B_{12}	$2\log_e B_{12}$	Evidence for M_1
<1	<0	Support for M_2
1–3	0–2	Barely worth mentioning
3–12	2–5	Positive
12–150	5–10	Strong
>150	>10	Very strong

where B_{12} is the Bayes factor of a comparison of models 1 and 2. Twice the log of the Bayes factor is on roughly the same scale as the more familiar likelihood-ratio test statistic. The Bayes factor of 3.9 for this analysis counts only as "positive" support for an African origin of the mtDNA.

Combining Data Across Studies

Since the Vigilant et al. (1991) study, many other studies have been performed

to infer the geographic area of the MRCA of modern humans. By and large, these studies have supported the idea that modern human mtDNA originated in Africa. How should new studies influence our belief about an African origin for modern human mtDNA?

In a Bayesian analysis, the posterior probabilities of hypotheses from a previous study can serve as the prior for the next study. For instance, the Vigilant et al. (1991) study, if done in a state of ignorance about the phylogenetic relationships of the sequences, would assign a probability of 0.80 to the idea that the origin of the sample was in Africa. This prior probability, as described above, considers all trees to be equally probable. The fact that 78 of the 135 human sequences were from people with origins in Africa implies that 0.8 of the trees will be reconstructed as having the MRCA in Africa.

The posterior probability of an origin in Africa for the modern human mtDNA, calculated in the previous section, describes how a person who started in a state of ignorance about the trees should change his or her opinion about an African origin for the sequences after observing the Vigilant et al. (1991) mitochondrial sequences. Subsequent studies can use the posterior probabilities from the Vigilant et al. (1991) data as the prior. This seems reasonable to do with phylogenetic trees, even though the root of one tree may not correspond to anything in another tree derived from another dataset because of nonoverlapping sample membership. We are interested in the geographic region of the root of the tree, not the root position per se.

To demonstrate this point, we analyzed 200 sequences, each 428 sites long, from the hypervariable region I (HVRI) of the mitochondrial control region from modern humans. We collected the sequences from the HvrBase database (Handt et al., 1998). We used the neandertal mitochondrial sequence as the outgroup (Krings et al., 1997). Sequences were sampled in such a way that 40 sequences came from each geographic region (Africa, Europe, the Americas, Asia, or Australia/Oceania). We performed a Bayesian analysis of the data under the GTR model with gamma-distributed rate variation. As before, the posterior probabilities of trees were approximated by using MCMC. We ran two chains (each with three heated and one cold chain) for 5×10^6 generations. The first 10^6 generations of the chains were discarded as the burn-in. The larger data set was more difficult to analyze than the original Vigilant et al. (1991) data, requiring a longer number of cycles to reach apparent stationarity. Hence, we ran the chain longer and discarded more of the initial cycles as the burn-in. We reconstructed the ancestral area of the MRCA of modern humans using the parsimony criterion. Because we had the same number of sequences from each geographic region of interest, the prior probability that the MRCA originated on continent i was simply 1 in 5. We confirmed this by generating a large sample of random trees and performing the parsimony reconstruction for the ancestral area on each. The prior and posterior probabilities of the MRCA of the sequences being in area i were

Area (i)	$f(M_i)$	$f(M_i \mid \boldsymbol{X})$
1 (Africa)	0.20	0.82
2 (Europe)	0.20	0.00
3 (Americas)	0.20	0.03
4 (Asia)	0.20	0.14
5 Australia/Oceania	0.20	0.01

These data, like the data of Vigilant et al. (1991), are consistent with an African origin for the modern human mitochondrion.

The initial analysis of the 201 mitochondrial sequences was performed under a uniform prior (on trees and on areas). What would the posterior probabilities of the five hypotheses be if the posterior probabilities from the Vigilant et al. (1991) data were used as the prior probabilities for the new data? The posterior probability of hypothesis M_i is

$$f'(M_i \mid \boldsymbol{X}) = \frac{f(M_i \mid \boldsymbol{X}) \times \frac{f'(M_i)}{f(M_i)}}{\sum_{j=1}^{5} f(M_j \mid \boldsymbol{X}) \times \frac{f'(M_j)}{f(M_j)}}$$

where $f(M_i)$ and $f(M_i \mid \boldsymbol{X})$ are the prior and posterior probabilities of the ith hypothesis used in the MCMC calculation, respectively, and $f'(M_i)$ is the prior probability of the ith hypothesis from the Vigilant et al. (1991) data. $f'(M_i \mid \boldsymbol{X})$ is the posterior probability of a geographic origin on

continent i for the mitochondrial sequences, given the new data from the Vigilant et al. (1991) prior. The posterior probability of a geographic origin in Africa for the new sequences from the Vigilant et al. (1991) prior is

$$f'(M_1 \mid \boldsymbol{X}) = \frac{0.82 \times \frac{0.94}{0.20}}{0.82 \times \frac{0.94}{0.20} + 0.00 \times \frac{0.01}{0.20} + 0.03 \times \frac{0.00}{0.20} + 0.14 \times \frac{0.03}{0.20} + 0.01 \times \frac{0.02}{0.20}} = 0.99$$

The prior and posterior probabilities for a geographic origin in area i become

Area (i)	$f'(M_i)$	$f'(M_i \mid \boldsymbol{X})$
1 (Africa)	0.94	0.99
2 (Europe)	0.01	0.00
3 (Americas)	0.00	0.00
4 (Asia)	0.03	0.01
5 Australia /Oceania	0.02	0.00

(Because of Monte Carlo error in the MCMC procedure, the small probabilities become zero when no trees are sampled that were consistent with the hypothesis.)

Conclusions

In a Bayesian analysis, parameters of a statistical model are treated as random variables with prior probability distributions. The general approach is to calculate the joint posterior probability for all the parameters. Inferences for any single parameter are then based on its marginal posterior probability. The posterior probability of a parameter is calculated by using Bayes's rule. In this study, we point out how Bayesian inference can be used to accommodate phylogenetic uncertainty when comparing five different models for the origin of modern human mtDNA. The analysis accommodates phylogenetic uncertainty by summing inferences over all possible phylogenetic trees, weighting each tree by its posterior probability. Moreover, the analysis allows the models to be compared through a comparison of the prior and posterior probabilities of the hypotheses. The original mitochondrial data collected by Vigilant et al. (1991) should cause a scientist who started off with a belief that all phylogenetic trees were equally probable to modify his or her belief about an African origin for the mtDNA from 0.80 to 0.94 after observing the data. The Bayes factor, or ratio of the marginal likelihoods, was 3.9, meaning that a person's belief in an African origin for the MRCA of the sequences changed by a factor of about 4 in favor of the hypothesis. A Bayes factor of 3 or 4 is typically considered positive, but not overwhelming, support for a hypothesis.

The same conclusion about the original Vigilant et al. (1991) study was made by Maddison et al. (1992), who pointed out that (1) among the large number of equally parsimonious trees, a large proportion were consistent with an origin of the mtDNA that was not in Africa, and (2) the root position of the tree was uncertain when the chimpanzee sequences were used as an outgroup. Their conclusion was that the Vigilant et al. (Maddison et al. 1992:122) data "do not unambiguously support an African origin of human mtDNA." Similarly, Penny et al. (1995) performed an extensive analysis of the landscape of phylogenetic trees for the Vigilant et al. (1991) data based on the parsimony criterion. They, too, found support for an African origin for the modern human mitochondrion. Their analysis involved performing many (>1,000) heuristic searches starting from different random trees. This procedure allowed them to find local optima in the landscape of phylogenetic trees. Importantly, these optima are consistent with an African origin for the sequences. The main disadvantage of the earlier studies was reliance on the parsimony method as an optimality criterion. The method outlined in this paper takes advantage of the strengths of likelihood-based optimality criteria and provides a basis for comparing different hypotheses about the trees. Importantly, the results of this study do not depend on any single tree being correctly estimated.

What has largely gone unappreciated is that the very decision to collect a certain number of DNA sequences from people with origins in Africa automatically induces a prior probability on the out-of-Africa hypothesis. It is useful, then, to compare the

prior probability of the hypothesis (before any sequences have been collected, but after the sample membership has been determined) with the probability of the hypothesis after collection of the data. We can certainly imagine situations in which the posterior probability of the hypothesis is high but indicates that the data support the alternative. For example, imagine that the posterior probability of an African origin of the sequences was 0.94, as it was in this study, but that the prior probability of the same hypothesis was 0.98. In this case, the data would actually support the hypothesis that the MRCA of the human mtDNA was not in Africa (in fact, the Bayes factor would be 0.32 in favor of the alternative hypothesis). This perspective suggests that the sampling strategy can determine how well a hypothesis can be tested. Although it is not clear what the optimal sampling strategy should be, it seems intuitive that it is easier to test a hypothesis that has a low prior probability than one with a high prior.

The analyses described in this paper can be modified in several ways. For one, this study used a uniform prior on all possible trees. However, this prior does not reflect the probability distribution of trees expected under a coalescence or birth–death process in which every labeled history has equal probability. Hence, one simple modification would be to perform the analyses with a coalescence prior. Another modification to the analyses would be to use a different method for reconstructing the ancestral geographic area of the human mtDNA. This study used parsimony, but one can imagine using a stochastic two-state model (Schluter, 1995; Schluter et al., 1997; Pagel, 1999; Schultz and Churchill, 1999) or even using a coalescence process with different populations connected by variable levels of migration (e.g., Beerli and Felsenstein, 1999). Finally, the trees could be reconstructed under a molecular clock constraint, which obviates the need for an outgroup because the molecular clock forces a root to the tree. These modifications could be incorporated into a Bayesian or maximum likelihood framework, with MCMC used to integrate over uncertainty in the nuisance parameters.

This paper examined only the data from one of the earliest molecular studies of the origin of modern humans and a new dataset collated from the HvrBase. Since the pioneering studies of Cann et al. (1987) and Vigilant et al. (1991), however, additional loci have been examined from much larger samples (Underhill et al., 2000). Moreover, fossil DNA sequences from neandertals and Australian aborigines have been collected (Krings et al., 1997; Adcock et al., 2001). These new datasets may provide a much more powerful test of the out-of-Africa hypothesis. Regardless of the data collected, careful consideration must be paid to how the data will modify beliefs about the out-of-Africa hypothesis. We hope that the framework discussed in this paper will clarify thinking about how molecular sequence data can be used to test biogeographic hypotheses when faced with phylogenetic uncertainty.

ACKNOWLEDGMENTS

This paper was improved through the suggestions of D. Penny, P. Lewis, C. Simon, and J. Thorne. This research was supported by National Science Foundation grants DEB-0075406 and MCB-0075404.

REFERENCES

ADCOCK, G. J., E. S. DENNIS, S. EASTEAL, G. A. HUTTLEY, L. S. JERMIN, W. J. PEACOCK, AND A. THORNE. 2001. Mitochondrial DNA sequences in ancient Australians: Implications for modern human origins. Proc. Nat. Acad. Sci. USA 98:537–542.

AWADALLA, P., A. EYRE-WALKER, AND J. M. SMITH. 1999. Linkage disequilibrium and recombination in Hominid mitochondrial DNA. Science 286:2524–2525.

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.

CANN, R. L., M. STONEKING, AND A. C. WILSON. 1987. Mitochondrial DNA and human evolution. Nature 325:31–36.

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

FITCH, W. M. 1971. Toward defining the course of evolution: Minimum change for a specific tree topology. Syst. Zool. 20:406–416.

GEYER, C. J. 1991. Markov chain Monte Carlo maximum likelihood. Pages 156–163 *in* Computing science and statistics: Proceedings of the 23rd Symposium on the Interface (E. Keramidas, ed.). Interface Foundation, Fairfax Station, Virginia.

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

HANDT, O., S. MEYER, AND A. VON HAESELER. 1998. Compilation of human mtDNA control region sequences. Nucleic Acids Res. 26:126–129.

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

HUELSENBECK, J. P., B. RANNALA, AND J. P. MASLY. 2000. Accommodating phylogenetic uncertainty in evolutionary studies. Science 288:2349–2350.

HUELSENBECK, J. P., AND F. RONQUIST. 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755

JEFFREYS, H. 1935. Some tests of significance, treated by the theory of probability. Proc. Cambridge Philos. Soc. 31:203–222.

JEFFREYS, H. 1961. Theory of probability. Oxford Univ. Press, Oxford.

KRINGS, M., A. STONE, R. W. SCHMITZ, H. KRAINITZKI, M. STONEKING, AND S. PÄÄBO. 1997. Neandertal DNA sequences and the origin of modern humans. Cell 90:19–30.

LARGET, B., AND D. SIMON. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16:750–759.

LAVINE, M., AND M. J. SCHERVISH. 1999. Bayes factors: What they are and what they are not. Am. Stat. 53:119–122.

LI, S. 1996. Phylogenetic tree construction using Markov chain Monte Carlo. Ph.D. dissertation, Ohio State Univ., Columbus.

MADDISON, D. R., M. RUVOLO, AND D. L. SWOFFORD. 1992. Geographic origins of human mitochondrial DNA: Phylogenetic evidence from control region sequences. Syst. Biol. 41:111–124.

MAU, B. 1996. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Ph.D. Dissertation, Univ. of Wisconsin, Madison.

MAU, B., AND M. NEWTON. 1997. Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. J. Comput. Graph. Stat. 6:122–131.

MAU, B., M. NEWTON, AND B. LARGET. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics 55:1–12.

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. Phys. 21:1087–1091.

NEWTON, M., B. MAU, AND B. LARGET. 1999. Markov chain Monte Carlo for the Bayesian analysis of evolutionary trees from aligned molecular sequences. *In* Statistics in molecular biology (F. Seillier-Moseiwitch, T. P. Speed, and M. Waterman, eds.). Monograph Series of the Institute of Mathematical Statistics.

PAGEL, M. 1999. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. Syst. Biol. 48:612–622.

PENNY, D., M. STEEL, P. J. WADDELL, AND M. D. HENDY. 1995. Improved analyses of human mtDNA sequences support a recent African origin for *Homo sapiens*. Mol. Biol. Evol. 12:863–882.

RAFTERY, A. 1995. Hypothesis testing and model selection. Pages 163–187 in Markov chain Monte Carlo in Practice (W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds.). Chapman and Hall, New York.

RANNALA, B., AND Z. YANG. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J. Mol. Evol. 43:304–311.

SCHLUTER, D. 1995. Uncertainty in ancient phylogenies. Nature 377:108–109.

SCHLUTER, D., T. PRICE, A. Ø. MOOERS, AND D. LUDWIG. 1997. Likelihood of ancestor states in adaptive radiation. Evolution 51:1699–1711.

SCHULTZ, T. R., AND G. A. CHURCHILL. 1999. The role of subjectivity in reconstructing ancestral character states: A Bayesian approach to unknown rates, states, and transformation asymmetries. Syst. Biol. 48:651–664.

STRINGER, C., AND R. MCKIE. 1996. African exodus. Jonathan Cape, London.

SWOFFORD, D. L. 1998. PAUP*: Phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, Massachusetts.

SWOFFORD, D., G. OLSEN, P. WADDELL, AND D. M. HILLIS. 1996. Phylogenetic inference. Pages 407–511 *in* Molecular systematics, 2nd edition (D. Hillis, C. Moritz, and B. Mable, eds.). Sinauer Associates, Sunderland, Massachusetts.

TAVARÉ, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lect. Math. Life Sci. 17:57–86.

TIERNEY, L. 1994. Markov chains for exploring posterior distributions (with discussion). Ann. Stat. 22:1701–1762.

UNDERHILL, P. A., P. SHEN, A. A. LIN, L. JIN, G. PASSARINO, W. H. YANG, E. KAUFFMAN, B. BONNÉ-TAMIR, J. BERTRANPETIT , P. FRANCALACCI, et al. 2000. Y chromosome sequence variation and the history of human populations. Nat. Genet. 26:358–361.

VIGILANT, L., M. STONEKING, H. HARPENDING, K. HAWKES, AND A. C. WILSON. 1991. African populations and the evolution of human mitochondrial DNA. Science 253:1503–1507.

WOLPOFF, M., AND R. CASPARI. 1997. Race and human evolution. Simon and Schuster, New York.

YANG, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396–1401.

YANG, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. 39:306–314.

YANG, Z., AND B. RANNALA. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. Mol. Biol. Evol. 14:717–724.

Received 14 February 2001; accepted 29 August 2001
Associate Editor: Jeff Thorne