Maximum Likelihood Implementation of an Isolation-with-Migration Model with Three Species for Testing Speciation with Gene Flow

Tianqi Zhu[1,2] and Ziheng Yang*[,3,4,5]

[1]School of Mathematical Sciences, Peking University, Beijing, China

[2]National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, 100190, Beijing, China

[3]Center for Computational and Evolutionary Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

[4]Computational Biochemistry Research Group, ETH Zurich, Zurich, Switzerland

[5]Department of Biology, Galton Laboratory, University College London, London, United Kingdom

***Corresponding author:** E-mail: z.yang@ucl.ac.uk.

Associate editor: Rasmus Nielsen

Abstract

We implement an isolation with migration model for three species, with migration occurring between two closely related species while an out-group species is used to provide further information concerning gene trees and model parameters. The model is implemented in the likelihood framework for analyzing multilocus genomic sequence alignments, with one sequence sampled from each of the three species. The prior distribution of gene tree topology and branch lengths at every locus is calculated using a Markov chain characterization of the genealogical process of coalescent and migration, which integrates over the histories of migration events analytically. The likelihood function is calculated by integrating over branch lengths in the gene trees (coalescent times) numerically. We analyze the model to study the gene tree-species tree mismatch probability and the time to the most recent common ancestor at a locus. The model is used to construct a likelihood ratio test (LRT) of speciation with gene flow. We conduct computer simulations to evaluate the LRT and found that the test is in general conservative, with the false positive rate well below the significance level. For the test to have substantial power, hundreds of loci are needed. Application of the test to a human–chimpanzee–gorilla genomic data set suggests gene flow around the time of speciation of the human and the chimpanzee.

Key words: coalescent, maximum likelihood, speciation, gene flow, isolation, migration.

Introduction

Genomic sequences from closely related species provide valuable information about the speciation process, such as the absence or presence of gene flow at the time of species formation (Patterson et al. 2006; Burgess and Yang 2008). However extracting such information requires probabilistic modeling of the genealogical relationships among sequences and powerful statistical inference methods that can take account of information and uncertainties from different sources, including gene tree-species tree conflicts due to ancestral polymorphism and lineage sorting, and uncertainties in the gene trees due to limited genetic variation at every locus, etc. In particular, the coalescent process is known to be highly variable and can create large fluctuations among loci or among genomic regions (Barton 2006).

Yang (2010) developed a likelihood ratio test (LRT) of speciation with gene flow, using genomic sequence data from three species (two closely related species 1 and 2 plus an out-group species 3), with one sequence sampled from each species. Gene flow at the time of split of species 1 and 2 is modeled as variation in the divergence time between species 1 and 2 across the genome, and the model is compared with a null model that assumes a constant divergence time. The model is implemented in the likelihood framework, with numerical integration used to average over coalescent times in the gene trees. Migration or gene flow is not explicitly considered in the model, which is thus only an approximate description of the genealogical process under the "isolation-with-migration" (IM) model.

IM models were implemented by Hey and Nielsen (2004) in the IM and IMa programs for two populations and have recently been extended to an arbitrary number of populations by Hey (2010a). Markov chain Monte Carlo (MCMC) algorithms are used to average over the gene tree topologies, the coalescent times, and the histories of migration events. The algorithm has to integrate over the number, directions, and times of migration events at every locus. While there are $s - 1$ coalescent times to integrate over in a gene tree for s sequences, there is no upper limit to the number of migration events, so that the space the MCMC algorithm has to sample from expands considerably, especially at high migration rates. As MCMC proposals altering migration histories without changing the gene tree topology or branch lengths will not affect the likelihood (i.e., the probability of the sequence data), the MCMC may be averaging over a huge nearly flat surface. As a result, MCMC algorithms for the IM model are computationally far more demanding than similar MCMC algorithms

under multispecies coalescent models without gene flow (Rannala and Yang 2003). For example, it is challenging to analyze data sets of 100 loci using the IM or IMa programs (see, e.g., table 1 in Hey 2010b), while data sets of 50,000 loci have been comfortably analyzed using the MCMCCOAL or BPP programs (Burgess and Yang 2008).

Thus in this era of ubiquitous genomics, there is a keen interest in fast computational algorithms suitable for analyzing a huge number of loci (10^4–10^5 loci, say) from only a few genomes (Wang and Hey 2010; Hobolth et al. 2011). In this paper, we report a maximum likelihood (ML) implementation of the IM model for three species. We focus on migration between the two closely related species 1 and 2, while species 3 is considered an out-group and does not exchange migrants with either species 1 or 2 or their common ancestor. This will be a proper implementation of the IM model for three species, compared with the approximation in Yang (2010). We use the Markov chain characterization of the genealogical process for the sample, which has been used in analysis of equilibrium population genetics models of migration, such as the stepping-stone model, the finite island model, or the general model specified using a migration matrix in the so-called structured coalescent framework (Notohara 1990; Wilkinson-Herbots 1998). Such models have been implemented in GENETREE (Bahlo and Griffiths 2000) and MIGRATE (Beerli and Felsenstein 1999, 2001; Beerli 2006), making it possible to estimate jointly the population size parameters (θs) and migration rates using genetic data. However, compared with those population genetics models, the IM model has the advantage of accommodating the relationships among the species/populations through the use of a phylogeny and may thus be more realistic for many real data sets. Furthermore, the IM model allows us to evaluate the role of gene flow during speciation.

A benefit of the Markov chain characterization of the genealogical process is that the probability density of gene tree topologies and branch lengths (coalescent times) can be calculated using the transition probability matrix for the Markov chain, $P(t)$, integrating over the histories of migration events analytically, leading to significant reduction in computation (Hobolth et al. 2011). As in Yang (2010), the integration over the branch lengths in the gene trees is achieved through numerical integration, and we discuss strategies to overcome the computational burden.

We use the IM model for three species to construct an LRT of speciation with gene flow, with the null hypothesis assuming no gene flow between species 1 and 2. In theory, the LRT can be conducted using data from species 1 and 2 only, under the two-species IM model studied by Wilkinson-Herbots (2008), Wang and Hey (2010), and Hobolth et al. (2011). However, inclusion of an out-group species adds valuable information in the form of the gene tree topology and branch lengths, and the gene tree-species tree conflict, thus increasing the power of the test. It should also make the test more robust to mutation rate variation among loci (Yang 1997, 2002). We then apply the test to a data set of genomic sequences from the human, chimpanzee, and gorilla, compiled by Burgess and Yang (2008).

Theory and Methods

An IM Model for Three Species and Its Markov Chain Characterization

Consider the species phylogeny for three species of figure 1*a*: ([1, 2], 3), where the two ancestral species are labelled 4 and 5. The data consist of multiple neutral loci, with one sequence sampled from each of the three species at every locus. As in Yang (2010), we assume no recombination within a locus and free recombination among loci. The gene trees describing the relationships among the three sampled sequences at any locus are depicted in figure 1*b*–*f*.

We consider migration between species 1 and 2, but assume no migration involving the out-group species 3. There are eight parameters in the model: θ_4, θ_5, τ_0, τ_1, θ_1, θ_2, M_{12}, and M_{21}. Here τ_0 and τ_1 are the two species divergence times, measured by the expected number of mutations per site, and $\theta_1 = 4N_1\mu$, $\theta_2 = 4N_2\mu$, $\theta_4 = 4N_4\mu$, and $\theta_5 = 4N_5\mu$ are the population size parameters for species 1, 2, 4, and 5, with the N's to be the effective population sizes and μ the mutation rate per site. If migration involving species 3 is allowed in the model or if two or more sequences are sampled from species 3, $\theta_3 = 4N_3\mu$ will have to be considered in the model as well. The migration rate M_{ij} is defined as the expected number of migrant individuals from population i to population j per generation: $M_{ij} = N_j m_{ij}$, where m_{ij} is the migration rate from populations i to j, defined as the proportion of individuals in population j that are immigrants from population i. With only one sequence from each of species 1 and 2, we expect the data to contain little information about the four parameters θ_1, θ_2, M_{12}, and M_{21}. Thus in this study, we assume that $\theta_1 = \theta_2 = \theta$, and $M_{12} = M_{21} = M$. As a result, six parameters are estimated from the data: $\theta = \{\theta_4, \theta_5, \tau_0, \tau_1, \theta, M\}$. This is referred to as the symmetric IM model for three species (SIM3s). Possible extensions to the model are discussed later.

When we trace the genealogy of the sample backward in time, the process can be described using three Markov chains for the three time epochs E_1, E_2, and E_3, defined by the species tree (fig. 1). Epoch E_1 goes from the present time to τ_1, when there exist three species: 1, 2, and 3. Epoch E_2 is from time τ_1 to τ_0, with two species: 3 and 5. Epoch E_3 goes from time τ_0 to infinity, with only one species: 4. During each time epoch, the genealogical process of coalescent and migration is described by the structured coalescent (Notohara 1990; Wilkinson-Herbots 1998).

Under the SIM3s model and given the data of one sequence from each of the three species, we have to consider only five states for the Markov chain during epoch E_1. These are 113, 123, 223, 13, and 23. Here, 113 means three sequences in the sample, with two from species 1 and the third from species 3; 123 means three sequences with one sequence from each of the three species; 13 means two sequences from species 1 and 3, and so on. During epoch E_2, there are only two species (3 and 5), and there are only two states in the Markov chain: 355 and 35, with the latter

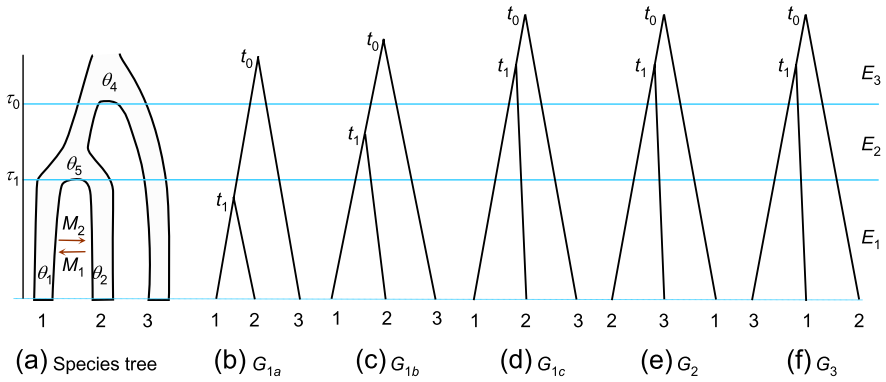

FIG. 1. (*a*) The species tree ([1, 2], 3) for three species, showing the parameters in the SIM3s model: θ_4, θ_5, τ_0, τ_1, $\theta_1 = \theta_2 = \theta$, and $M_{21} = M_{12} = M$. As no migration involving species 3 is assumed in the model, and as one sequence is sampled from each species, θ_3 for species 3 is not a parameter in the model. The five possible gene trees for any locus are shown in b-f. If sequences 1 and 2 coalesce in species 1 or 2, the resulting gene tree will be G_{1a} (*b*), and if they coalesce in the common ancestor of species 1 and 2, the resulting gene tree will be G_{1b} (*c*). Otherwise three gene trees G_{1c}, G_2, and G_3 are possible as shown in (*d–f*). Gene trees G_{1a}, G_{1b}, and G_{1c} have the same tree topology as the species tree, while G_2 and G_3 have different topologies from the species tree. Coalescent times are represented by node ages t_0 and t_1 in each gene tree.

meaning that the sequences from species 1 and 2 have coalesced. Note that when the process leaves epoch E_1 to enter E_2, the sequences and the states may be relabelled. For example, state 113 in epoch E_1 becomes state 355 in epoch E_2. Epoch E_3 has only one species (4), and the genealogical process is the simple single-population coalescent.

Consider the genealogical process in any population i and suppose there are n_i ancestral lineages in the population. If time is measured in generations, coalescent occurs in population i at the rate of $n_i(n_i - 1)/2 \times 1/(2N_i)$, while "backward" migration occurs from population i to population j at the rate $n_i m_{ji}$. Here, the $i \to j$ migration is "backward" due to the coalescent worldview in which time runs backward and means migration from populations j to i in the real world. Divide both coalescent and migration rates by μ, so that time is measured by the expected number of mutations per site. Then, the coalescent rate becomes $n_i(n_i - 1)/2 \times 2/\theta_i$, while the (real-world) migration rate from population j into i becomes $n_i m_{ji}/\mu = n_i \times 4M_{ji}/\theta_i$. Thus the rate matrix (generator) for the Markov chain can be constructed: $Q^{(1)}$ and $Q^{(2)}$ for epochs E_1 and E_2, respectively.

For the special model SIM3s considered in this paper, the genealogical process in epoch E_1 concerns really two populations (species 1 and 2), and the transition probability matrix $P^{(1)}(t) = \exp(Q^{(1)}t)$ is analytically tractable. Let the time unit be one expected mutation per site, and define the coalescent and migration rates as $c = 2/\theta$ and $w = 4M/\theta$, to simplify the formulae. The rate matrix $Q^{(1)}$ is then

	113	123	223	13	23
113	$-(2w + c)$	$2w$	0	c	0
123	w	$-2w$	w	0	0
223	0	$2w$	$-(2w + c)$	0	c
13	0	0	0	$-w$	w
23	0	0	0	w	$-w$

(1)

This is a special case of the matrix in equation (1) of Hobolth et al. (2011) for the IM model for two species. This $Q^{(1)}$ matrix has the eigenvalues $\lambda_1 = 0$, $\lambda_2 = -(2w + c)$, $\lambda_3 = -2w$, $\lambda_{4,5} = -\frac{1}{2}\left(4w + c \pm \sqrt{16w^2 + c^2}\right)$, all of which are distinct as $w > 0$ and $c > 0$. The transition probability matrix can be calculated analytically through the spectral decomposition of $Q^{(1)}$ (see Appendix).

The Probability Distribution of Gene Genealogies

The Markov chain formulation allows us to calculate the probability distribution of the gene trees and branch lengths under the SIM3s model. At any locus, there are five possible gene trees, shown in figure 1, with their coalescent times t_1 and t_0. Gene trees G_{1a}, G_{1b}, and G_{1c} have the same topology as the species tree, while G_2 and G_3 have different topologies. G_{1a} is the result of sequences from species 1 and 2 coalescing in epoch E_1; this is possible because migrations between species 1 and 2 can bring the two sequences into the same population (species 1 or 2). G_{1b} results from sequences from species 1 and 2 coalescing in epoch E_2 (in species 5). If no coalescent occurs in epochs E_1 or E_2, gene trees G_{1c}, G_2, and G_3 will be generated, each with the same probability. Below we derive the probability densities of the gene tree and branch lengths (coalescent times) under the SIM3s model for each of the gene trees G_{1a}, G_{1b}, G_{1c}, G_2, and G_3.

For gene tree G_{1a}, a coalescent event occurs in epoch E_1 so that right before the coalescent event, the state of the Markov chain must be either 113 or 223. Thus

$$\begin{aligned} f(G_{1a}, t_0, t_1) &= \left(P^{(1)}_{123,113}(t_1)\frac{2}{\theta} + P^{(1)}_{123,223}(t_1)\frac{2}{\theta}\right) \times \\ &\quad P^{(2)}_{35,35}(\tau_0 - \tau_1)\frac{2}{\theta_4}e^{-2(t_0-\tau_0)/\theta_4}, \qquad (2) \\ &= f(G_{1a}, t_1)\frac{2}{\theta_4}e^{-2(t_0-\tau_0)/\theta_4}, \end{aligned}$$

where $t_0 > \tau_0$ and $0 < t_1 < \tau_1$. The joint density of G_{1a} and t_1 can be calculated using equation (27) in the Appendix as

$$\begin{aligned} f(G_{1a}, t_1) &= P^{(1)}_{123,113}(t_1)\frac{2}{\theta} + P^{(1)}_{123,223}(t_1)\frac{2}{\theta} = 2P^{(1)}_{123,113}(t_1) \times \frac{2}{\theta} \\ &= \frac{2w}{a}e^{-\frac{1}{2}(c+4w-a)t_1}\left(1 - e^{-at_1}\right) \times c \qquad (3) \\ &= \frac{8M}{\theta\sqrt{64M^2+1}}e^{-\frac{t_1}{\theta}\left(8M+1-\sqrt{64M^2+1}\right)}\left(1 - e^{\frac{2t_1}{\theta}\sqrt{64M^2+1}}\right), \end{aligned}$$

where $a = \sqrt{c^2 + 16w^2}$. Note that as we assume no migration between species 3 and 5, $P^{(2)}_{35,35}(\tau_0 - \tau_1) = 1$.

For gene tree G_{1b}, no coalescent occurs in epoch E_1 but one occurs in E_2, at time t_1:

$$\begin{aligned} f(G_{1b}, t_0, t_1) &= f(G_{1b}, t_1) f(t_0 | G_{1b}, t_1) \\ &= \left(\sum_{s_1 \in A_3} P^{(1)}_{123,s_1}(\tau_1) \right) \times \left(P^{(2)}_{355,355}(t_1 - \tau_1) \frac{2}{\theta_5} \right) \times \\ &\quad P^{(2)}_{35,35}(\tau_0 - t_1) \times \frac{2}{\theta_4} e^{-2(t_0 - \tau_0)/\theta_4} \\ &= (1 - P(G_{1a})) \times \frac{2}{\theta_5} e^{-2(t_1 - \tau_1)/\theta_5} \times \frac{2}{\theta_4} e^{-2(t_0 - \tau_0)/\theta_4}, \end{aligned} \tag{4}$$

with $t_0 > \tau_0$ and $\tau_1 < t_1 < \tau_0$. Here $A_3 = \{113, 123, 223\}$ is the set of states with three sequences in epoch E_1. $P(G_{1a})$ is the probability of gene tree G_{1a}, that is, the probability that sequences 1 and 2 coalesce before τ_1:

$$\begin{aligned} P(G_{1a}) &= 1 - \sum_{s_1 \in A_3} P^{(1)}_{123,s_1}(\tau_1) \\ &= 1 - \frac{1}{2a} \left[(c + 4w + a) e^{-\frac{\tau_1}{2}(c + 4w - a)} \right. \\ &\quad \left. - (c + 4w - a) e^{-\frac{\tau_1}{2}(c + 4w + a)} \right] \\ &= 1 - \left[\left(\frac{8M + 1}{2\sqrt{64M^2 + 1}} + \frac{1}{2} \right) e^{-\frac{\tau_1}{\theta}(8M + 1 - \sqrt{64M^2 + 1})} \right. \\ &\quad \left. - \left(\frac{8M + 1}{2\sqrt{64M^2 + 1}} - \frac{1}{2} \right) e^{-\frac{\tau_1}{\theta}(8M + 1 + \sqrt{64M^2 + 1})} \right]. \end{aligned} \tag{5}$$

All the three states in A_3 become 355 when the genealogical process enters epoch E_2. Also $P^{(2)}_{35,35}(\tau_0 - t_1) = 1$ because of our assumption of no migration between species 3 and 5.

The densities $f(G_{1a}, t_1)$ and $f(G_{1b}, t_1)$ were obtained earlier using a different approach by Wilkinson-Herbots (2008), who studied the distribution of the coalescent time t between two sequences in an IM model where a panmictic ancestral population gave rise to n populations time τ ago. The density $f(G_{1a}, t_1)$ in equation (3) is Wilkinson–Herbots's equation (19), for $t \leq \tau$ (or $t_1 < \tau_1$ in our notation). This is also equation (3.6) of Nath and Griffiths (1993), truncated at $t_1 \leq \tau_1$. The density $f(G_{1b}, t_1)$ in equation (4) is Wilkinson–Herbots's equation (19) for $t > \tau$, truncated at τ_0.

Finally for gene trees G_{1c}, G_2, and G_3, with times $t_0, t_1 > \tau_0$,

$$\begin{aligned} f(G_k, t_0, t_1) &= \frac{1}{3} \left(\sum_{s_1 \in A_3} P^{(1)}_{123,s_1}(\tau_1) \right) P^{(2)}_{355,355}(\tau_0 - \tau_1) \times \\ &\quad \frac{6}{\theta_4} e^{-6(t_1 - \tau_0)/\theta_4} \frac{2}{\theta_4} e^{-2(t_0 - t_1)/\theta_4} \\ &= \frac{1}{3}(1 - P(G_{1a})) \times e^{-2(\tau_0 - \tau_1)/\theta_5} \times \\ &\quad \frac{6}{\theta_4} e^{-6(t_1 - \tau_0)/\theta_4} \frac{2}{\theta_4} e^{-2(t_0 - t_1)/\theta_4}, \end{aligned} \tag{6}$$

with $k = 1c$, 2, or 3. Here, $A_3 = \{113, 123, 223\}$ is again the set of states with three sequences ancestral to the sample in epoch E_1 and all those states become 355 in epoch E_2. The sum in equation (6) is over all paths in which there is no coalescent in epochs E_1 or E_2. In epoch E_3, two of the three sequences coalesce with waiting time $t_1 - \tau_0$, and then the second coalescent event occurs with waiting time $t_0 - t_1$: those waiting times have independent exponential distributions, with rates $6/\theta_4$ and $2/\theta_4$, respectively. Finally, $P^{(2)}_{355,355}(\tau_0 - \tau_1) = e^{-2(\tau_0 - \tau_1)/\theta_5}$ is the probability that the two sequences in species 5 do not coalesce during the time interval $\tau_0 - \tau_1$ over epoch E_2.

Equations (2), (4), and (6) provide a full characterization of gene trees under the SIM3s model. Those densities will be used in the calculation of the likelihood function.

Calculation of the Likelihood Function

We assume that the three sequences at every locus are already aligned, with alignment gaps removed. We assume the JC69 mutation model (Jukes and Cantor 1969) to correct for multiple hits. The different loci are assumed to have the same mutation rate, although it is straightforward to incorporate relative locus rates if these are externally estimated, for example, by using an out-group species (Yang 2002). The data at any site in the alignment will fall into five possible categories, called site patterns: *xxx*, *xxy*, *yxx*, *xyx*, and *xyz*, where *x*, *y*, and *z* are any distinct nucleotides. The data at any locus i are summarized as the counts of sites with those patterns: $D_i = \{n_{i0}, n_{i1}, n_{i2}, n_{i3}, n_{i4}\}$, and $D = \{D_i\}$ represents the data at all L loci. The likelihood calculation is quite similar to that in Yang (2002, 2010). The probability of data at locus i is

$$f(D_i; \boldsymbol{\theta}) = \sum_{k \in \{1a, 1b, 1c, 2, 3\}} \iint P(D_i | G_k, t_0, t_1) f(G_k, t_0, t_1) \mathrm{d}t_0 \mathrm{d}t_1. \tag{7}$$

The log likelihood is a sum over loci

$$\ell(\boldsymbol{\theta}; D) = \sum_{i=1}^{L} \log f(D_i; \boldsymbol{\theta}). \tag{8}$$

The probability of data at locus i given the gene tree G_k and coalescent times or node ages t_0 and t_1 (see fig. 1) is given by the multinomial probabilities

$$\begin{aligned} P(D_i | G_k, t_0, t_1) &= C \times p_0^{n_{i0}} p_1^{n_{i1}} p_2^{n_{i2} + n_{i3}} p_4^{n_{i4}}, \quad k = 1a, 1b, 1c, \\ P(D_i | G_2, t_0, t_1) &= C \times p_0^{n_{i0}} p_1^{n_{i2}} p_2^{n_{i3} + n_{i1}} p_4^{n_{i4}}, \\ P(D_i | G_3, t_0, t_1) &= C \times p_0^{n_{i0}} p_1^{n_{i3}} p_2^{n_{i1} + n_{i2}} p_4^{n_{i4}}, \end{aligned} \tag{9}$$

where

$$\begin{aligned} p_0(t_0, t_1) &= \mathrm{prob}(xxx) = \left(1 + 3e^{-8t_1/3} + 6e^{-8t_0/3} \right. \\ &\quad \left. + 6e^{-(8t_0 + 4t_1)/3}\right)/16, \\ p_1(t_0, t_1) &= \mathrm{prob}(xxy) = \left(3 + 9e^{-8t_1/3} - 6e^{-8t_0/3} \right. \\ &\quad \left. - 6e^{-(8t_0 + 4t_1)/3}\right)/16, \\ p_2(t_0, t_1) &= \mathrm{prob}(yxx) = \left(3 - 3e^{-8t_1/3} + 6e^{-8t_0/3} \right. \\ &\quad \left. - 6e^{-(8t_0 + 4t_1)/3}\right)/16, \\ p_3(t_0, t_1) &= p_2(t_0, t_1), \\ p_4(t_0, t_1) &= \mathrm{prob}(xyz) = \left(6 - 6e^{-8t_1/3} - 12e^{-8t_0/3} \right. \\ &\quad \left. + 12e^{-(8t_0 + 4t_1)/3}\right)/16 \end{aligned} \tag{10}$$

(Yang 1994). Thus, equation (7) becomes

$$f(D_i;\boldsymbol{\theta}) = \int_{\tau_0}^{\infty}\int_{0}^{\tau_1} P(D_i|G_{1a},t_0,t_1) \times 2P^{(1)}_{123,113}(t_1)\frac{2}{\theta} \times \frac{2}{\theta_4}e^{-\frac{2}{\theta_4}(t_0-\tau_0)}dt_1dt_0$$
$$+ \int_{\tau_0}^{\infty}\int_{\tau_1}^{\tau_0} P(D_i|G_{1b},t_0,t_1)(1-P(G_{1a}))\frac{2}{\theta_5}e^{-\frac{2}{\theta_5}(t_1-\tau_1)}\frac{2}{\theta_4}e^{-\frac{2}{\theta_4}(t_0-\tau_0)}dt_1dt_0 \quad (11)$$
$$+ \frac{1}{3}(1-P(G_{1a}))e^{-\frac{2}{\theta_5}(\tau_0-\tau_1)}\int_{\tau_0}^{\infty}\int_{\tau_0}^{t_0}\left[\sum_{k\in\{1c,2,3\}} P(D_i|G_k,t_0,t_1)\right] \times \frac{6}{\theta_4}e^{-\frac{6}{\theta_4}(t_1-\tau_0)}\frac{2}{\theta_4}e^{-\frac{2}{\theta_4}(t_0-t_1)}dt_1dt_0$$

$$= \int_0^{\infty}\int_0^{2\tau_1/\theta} P(D_i|G_{1a},\tau_0+\frac{1}{2}\theta_4x_0,\frac{1}{2}\theta x_1) \times 2P^{(1)}_{123,113}\left(\frac{1}{2}\theta x_1\right) \times e^{-x_0}dx_1dx_0$$
$$+ (1-P(G_{1a}))\int_0^{\infty}\int_0^{\frac{2}{\theta_5}(\tau_0-\tau_1)} P\left(D_i|G_{1b},\tau_0+\frac{1}{2}\theta_4x_0,\tau_1+\frac{1}{2}\theta_5x_1\right) \times e^{-x_1}e^{-x_0}dx_1dx_0$$
$$+ (1-P(G_{1a}))e^{-\frac{2}{\theta_5}(\tau_0-\tau_1)} \times \int_0^{\infty}\int_0^{\infty}\left[\sum_{k\in\{1c,2,3\}} P\left(D_i|G_k,\tau_0+\frac{1}{2}\theta_4x_0+\frac{1}{2}\theta_4x_1,\tau_0+\frac{1}{2}\theta_4x_1\right)\right] \times e^{-3x_1}e^{-x_0}dx_1dx_0.$$

The transform from t_0 and t_1 to new variables x_0 and x_1 is applied to increase the efficiency of numerical integration, by causing the quadrature points to be distributed in regions where the integrand is high (see below).

Nonidentifiability of the SIM3s Model

A statistical model is said to be nonidentifiable if there exist two sets of parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, for which the data have the same probability, that is,

$$f(D;\boldsymbol{\theta}) = f(D;\boldsymbol{\theta}'), \text{ for all } D. \quad (12)$$

Nonidentifiable models are often due to errors in model formulation and should be avoided (see, e.g., Rannala 2002). For the SIM3s model considered in this paper, we have

$$f(D_i|G_k,t_0,t_1,\boldsymbol{\theta}) = f(D_i|G_k,t_0,t_1). \quad (13)$$

Given the gene tree topology G_k and branch lengths (node ages t_0 and t_1), the probability of the sequence data does not depend on the parameters $\boldsymbol{\theta}$. Thus if there exist $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ that produce the same probability distributions of gene tree topologies and branch lengths, that is, if

$$f(G_k,t_0,t_1;\boldsymbol{\theta}) = f(G_k,t_0,t_1;\boldsymbol{\theta}'), \text{ for all } G_k,\ t_0,\ \text{and } t_1, \quad (14)$$

the model will be nonidentifiable.

From equations (2), (4), and (6), it is clear that for equation (14) to hold for all t_0 and t_1, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ must have the same θ_4 and θ_5. We will, in addition, fix τ_0 and τ_1 and demonstrate that different values of θ and M in the SIM3s model can produce the same $f(G_k, t_0, t_1; \boldsymbol{\theta})$ for all G_k, t_0, t_1. It is easy to see that $f(G_k, t_0, t_1; \theta)$, for k = 1a, 1b, 1c, 2, 3, are functions of a and b, defined as

$$a = \sqrt{16w^2+c^2} = \frac{2}{\theta}\sqrt{64M^2+1},$$
$$b = 4w + c = \frac{2(8M+1)}{\theta}. \quad (15)$$

If different values of c and w (or θ and M) correspond to the same values of a and b, the model will be nonidentifiable. From equation (15), w satisfies the following quadratic equation

$$32w^2 - 8bw + b^2 - a^2 = 0. \quad (16)$$

As $\Delta = 64(2a^2 - b^2) = 64(4w - c)^2 \geq 0$, the equation always has two valid roots, which may be identical. Thus with the four parameters (τ_0, τ_1, θ_4, θ_5) fixed, the points (c, w) and (c^*, w^*) = ($4w$, $c/4$), or the points (θ, M) and $(\theta^*, M^*) = (\frac{\theta}{8M}, \frac{1}{64M})$ always correspond to the same a and b, and thus always have the same log likelihood. If $(\hat{\theta}, \hat{M})$ is a local optimum, $(\hat{\theta}^*, \hat{M}^*) = (\frac{\hat{\theta}}{8\hat{M}}, \frac{1}{64\hat{M}})$ will also be a local optimum, with the same log likelihood. If $\hat{M} = 1/8$, the two points coincide.

The nonidentifiability means that it is not possible to estimate θ and M under the SIM3s model using multiloci data of three sequences, with one sequence from each species. Nevertheless, the LRT is still valid, as the same number of parameters are involved in the model if we use (a, b) as parameters instead of (θ, M).

Numerical Integration, ML Estimation, and Implementation

Each evaluation of the log likelihood function (8 and 11) requires calculation of 3L 2-D integrals for data of L loci. Following Yang (2010), we use numerical integration to calculate them. Gaussian quadrature is used, with 2-D integrals calculated by an iterated use of the 1-D algorithm (the so-called product rule). The computation is proportional to K^2. To decide how large K should be, we analyzed five simulated data sets with L = 15,000 loci, using K = 16 and 32, and found they produced similar results, with the difference in $\Delta\ell$ to be less than

0.01, suggesting that $K = 16$ is large enough. This value is used in calculations of this paper.

Note that the probabilities $P(D_i|G_k, t_0, t_1)$ of equations (10) and (11) are very small and vary over many orders of magnitude depending on t_0 and t_1. To avoid underflows and overflows, the highest log likelihood at the locus, ℓ_{max}, calculated at the ML tree topology and branch lengths, is used for scaling: the integrand of equation (11) are divided by $e^{\ell_{max}}$ before the terms are summed up.

We note that the eigenvalues and eigenvectors of the rate matrix $Q^{(1)}$ do not change over loci, neither do the site-pattern probabilities p_0–p_4 of equation (10). Those quantities are thus calculated prior to the numerical integration to save computation.

The SIM3s model is implemented in the C program 3s, using the optimization routine in the PAML package (Yang 2007) to find the ML estimates (MLEs) numerically. To validate the program, we simulated large data sets of $L = 10^5$ or 5×10^5 loci under the SIM3s model (Zhang et al. 2011). The MLEs of parameters under the same model (including θ and M) are found to be very close to the true values, although for θ and M, another set of values also have the same log likelihood. We did simple benchmarking using the hominoid data sets (see below) on a Windows laptop with an intel i7 CPU at 2.67 GHz. For the X-chromosome (or autosomal) data sets of $L = 510$ (or 9,861) loci, the likelihood iteration under SIM3s takes ~1 min (~20 min), with ~91% (or ~99.7%) of the computation spent on calculation of the probabilities of the data for different branch lengths (eq. 9). The log likelihood improves quickly during the initial stage of the iteration, but becomes slow when it is close to the optimum: roughly it spent ~90% of time after it is within 0.1 log-likelihood units of the optimum. This is partly because of the strong correlation in estimates of θ and M under the model (see below). The program is written in ANSI C, and can be compiled for different platforms. It is available at http://abacus.gene.ucl.ac.uk/software/.

Further Characterizations of the SIM3s Model

Here, we examine some predictions of the SIM3s model concerning the gene trees, such as the gene-tree height t_0 when one sequence is sampled from each of the three species, and the divergence time t_{12} between a pair of sequences from species 1 and 2.

Besides equation (5) giving the probability of gene tree G_{1a}, the probabilities for the other gene trees can be derived as

$$\begin{aligned} P(G_{1b}) &= (1 - P(G_{1a}))\left(1 - e^{-2(\tau_0-\tau_1)/\theta_5}\right), \\ P(G_k) &= \tfrac{1}{3}\left[1 - P(G_{1a}) - P(G_{1b})\right] \\ &= \tfrac{1}{3}\left[1 - P(G_{1a})\right]e^{-2(\tau_0-\tau_1)/\theta_5},\ k = 1c, 2, 3. \end{aligned} \tag{17}$$

The "gene tree-species tree mismatch probability" is the probability that the gene tree for any locus differs from the species tree (Takahata et al. 1995; Yang 2002)

$$\begin{aligned} P_{SG} &= P(G_2) + P(G_3) \\ &= \frac{2}{3}[1 - P(G_{1a}) - P(G_{1b})] \\ &= \frac{1}{3a}[(4w + c + a)e^{a\tau_1} - (4w + c - a)]e^{-\frac{\tau_1}{2}(c+4w+a)} \\ &\quad \times e^{-\frac{2}{\theta_5}(\tau_0-\tau_1)} \\ &= \frac{2}{3}\left[\left(\frac{8M+1}{2\sqrt{64M^2+1}} + \frac{1}{2}\right)e^{-\frac{\tau_1}{\theta}(8M+1-\sqrt{64M^2+1})}\right. \\ &\quad \left. - \left(\frac{8M+1}{2\sqrt{64M^2+1}} - \frac{1}{2}\right)e^{-\frac{\tau_1}{\theta}(8M+1+\sqrt{64M^2+1})}\right]e^{-\frac{2}{\theta_5}(\tau_0-\tau_1)}. \end{aligned} \tag{18}$$

If $M = 0$, we have $P(G_{1a}) = 0$, and

$$P_{SG} = \frac{2}{3}e^{-\frac{2}{\theta_5}(\tau_0-\tau_1)}, \tag{19}$$

as given by Hudson (1983). If $M \to \infty$, species 1 and 2 will be one population with size 2θ, so that $P(G_{1a}) \to 1 - e^{-\tau_1/\theta}$ and

$$P_{SG} \to \frac{2}{3}e^{-\frac{1}{\theta}\tau_1}e^{-\frac{2}{\theta_5}(\tau_0-\tau_1)}. \tag{20}$$

P_{SG} decreases monotonically when M increases from 0 to ∞. Figures 2*a* and *d* show the probabilities of gene trees for two sets of parameter values. Note that $P_{SG} = 2P(G_{1c})$.

The probability density of the time to the most recent common ancestor or gene tree height (t_0) can be derived by conditioning on the gene tree

$$\begin{aligned} f(t_0) &= \sum_{k\in(1a,1b,1c,2,3)} P(G_k)f(t_0|G_k) \\ &= [P(G_{1a}) + P(G_{1b})]\frac{2}{\theta_4}e^{-2(t_0-\tau_0)/\theta_4} + 3P(G_{1c}) \times \\ &\quad \frac{3}{\theta_4}\left(e^{-2(t_0-\tau_0)/\theta_4} - e^{-6(t_0-\tau_0)/\theta_4}\right) \\ &= (1 - [1 - P(G_{1a})]e^{-2(\tau_0-\tau_1)/\theta_5}) \times \\ &\quad \frac{2}{\theta_4}e^{-2(t_0-\tau_0)/\theta_4} + [1 - P(G_{1a})]e^{-2(\tau_0-\tau_1)/\theta_5} \times \\ &\quad \frac{3}{\theta_4}\left(e^{-2(t_0-\tau_0)/\theta_4} - e^{-6(t_0-\tau_0)/\theta_4}\right), \end{aligned} \tag{21}$$

for $t_0 > \tau_0$. Note that in the case of gene trees G_{1c}, G_2, and G_3, the density of t_0 is given as the sum (or convolution) of two independent exponential variables with means $\theta_4/6$ and $\theta_4/2$ as $\frac{3}{\theta_4}\left(e^{-2(t_0-\tau_0)/\theta_4} - e^{-6(t_0-\tau_0)/\theta_4}\right)$.

The density (eq. 21) for two sets of parameter values are shown in figure 2*b* and *e*. The expectation can be obtained as

$$\begin{aligned} E(t_0) &= E(E(t_0|G_k)) \\ &= [1 - (1 - P(G_{1a}))e^{-2(\tau_0-\tau_1)/\theta_5}] \times \left(\tau_0 + \frac{\theta_4}{2}\right) \\ &\quad + (1 - P(G_{1a}))e^{-2(\tau_0-\tau_1)/\theta_5} \times \left(\tau_0 + \frac{2\theta_4}{3}\right) \\ &= \tau_0 + \frac{\theta_4}{2}\left[1 + \frac{1}{3}(1 - P(G_{1a}))e^{-2(\tau_0-\tau_1)/\theta_5}\right]. \end{aligned} \tag{22}$$

Next, we study the divergence time t_{12} between two sequences sampled from species 1 and 2. The density can be derived by conditioning on the gene trees

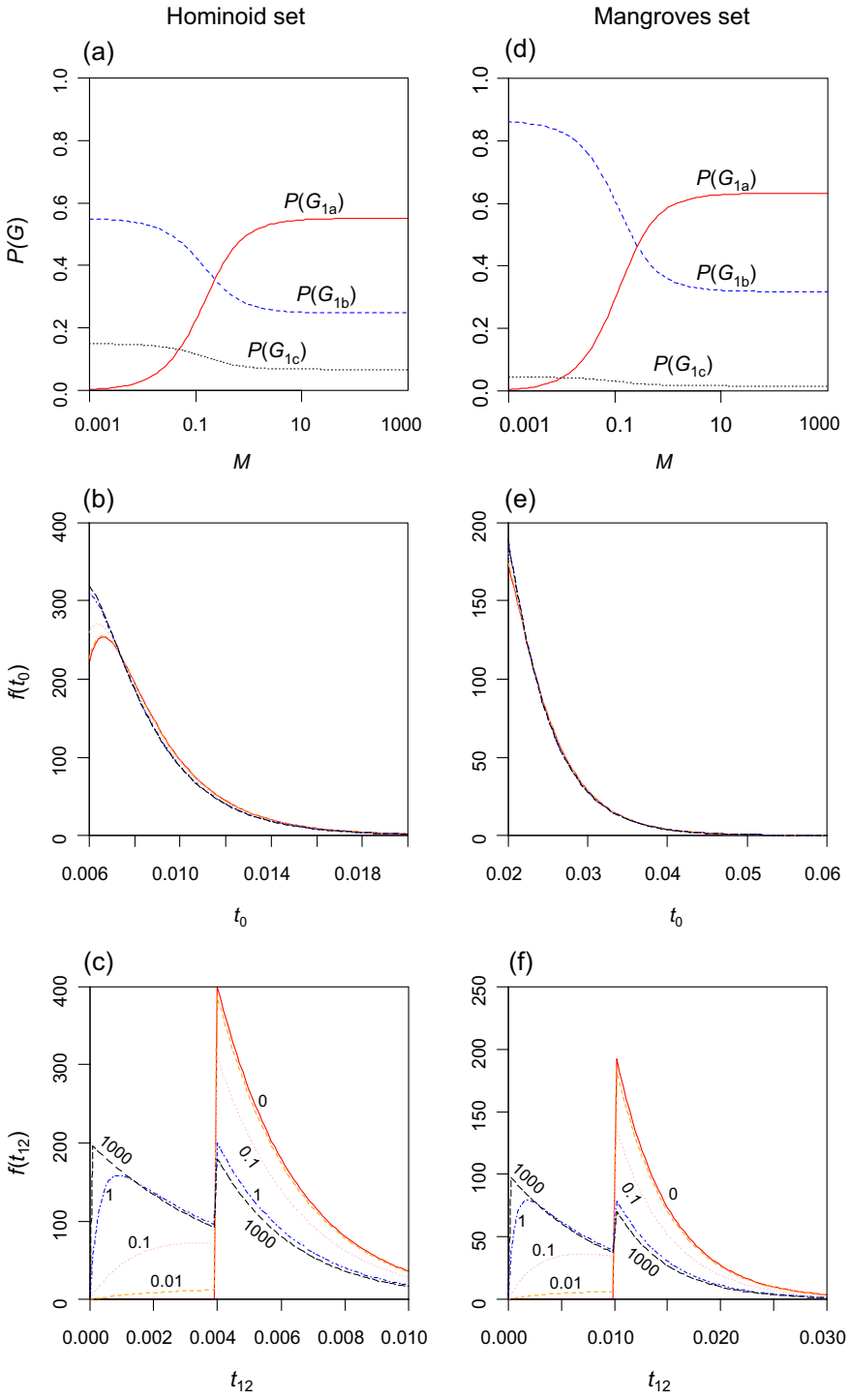

Fig. 2. Characterization of gene trees at different migration rates M under the SIM3s model. Panels *a* and *d* show the probabilities of gene trees G_{1a}, G_{1b}, and G_{1c}, with $P(G_{1a}) + P(G_{1b}) + 3P(G_{1c}) = 1$, and with $P_{SG} = 2P(G_{1c})$ to be the species tree-gene tree mismatch probability. Panels *b* and *e* show the density function of t_0 (tree height) for three sequences, one each from the three species. The five curves from top to bottom are for $M = 0, 0.01, 0.1, 1, 1000$. Panels *c* and *f* show the density function of t_{12}, the divergence time between two sequences sampled from species 1 and 2. The five curves are for $M = 0, 0.01, 0.1, 1, 1000$. Note that when $M = 0$, $f(t_{12}) = 0$ for $t_{12} < \tau_1$. The two sets of parameter values are $\theta_4 = \theta_5 = \theta = 0.005$, $\tau_0 = 0.006$, $\tau_1 = 0.004$ for the hominoid set (*a*–*c*); and $\theta_4 = \theta_5 = \theta = 0.01$, $\tau_0 = 0.02$, $\tau_1 = 0.01$ for the mangroves set (*d*–*f*).

Table 1. The False Positive Rate, Percentage of Zeros, and 95% Quantile of the Null Distribution of the LRT Statistic ($2\Delta\ell$).

Parameters	$L = 10$			100			1,000			15,000		
Set 1	0.000	0.84	0.28	0.001	0.66	2.30	0.005	0.59	2.18	0.008	0.52	2.88
Set 2	0.002	0.85	1.28	0.002	0.78	1.30	0.004	0.70	2.20	0.005	0.61	2.24
Set 3	0.000	0.93	0.42	0.002	0.90	0.64	0.005	0.88	0.64	0.012	0.88	0.82
Set 4	0.002	0.83	1.56	0.006	0.74	2.10	0.006	0.69	1.88	0.005	0.66	1.90

NOTE.—In each cell, the three numbers are 1) the false positive rate when the test is conducted using $\chi^2_{2,5\%}=5.99$, 2) the proportion of replicates in which the test statistic $2\Delta\ell=0$, and 3) the 95% quantile of the null distribution (as opposed to 5.99). Sequence length is 500 bp at each of the L loci. The number of replicates is 1,000. Data are simulated using MCCOAL under model M0 using four sets of parameter values, as follows: Set 1 (hominoid): $\theta_4 = \theta_5 = 0.005$, $\tau_0 = 0.006$, $\tau_1 = 0.004$; Set 2 (mangroves): $\theta_4 = \theta_5 = 0.01$, $\tau_0 = 0.02$, $\tau_1 = 0.01$; Set 3: $\theta_4 = 0.02$, $\theta_5 = 0.03$, $\tau_0 = 0.06$, $\tau_1 = 0.04$; Set 4: $\theta_4 = 0.02$, $\theta_5 = 0.01$, $\tau_0 = 0.02$, $\tau_1 = 0.01$.

$$f(t_{12}) = \begin{cases} f(G_{1a}, t_{12}) \quad (\text{see equation 8}), & \text{if } 0<t_{12}<\tau_1, \\ (1 - P(G_{1a}))\frac{2}{\theta_5}e^{-2(t_{12}-\tau_1)/\theta_5}, & \text{if } \tau_1<t_{12}<\tau_0, \\ (1 - P(G_{1a}))e^{-2(\tau_0-\tau_1)/\theta_5} \times \frac{2}{\theta_4}e^{-2(t_{12}-\tau_0)/\theta_4}, & \text{if } t_{12}>\tau_0. \end{cases} \quad (23)$$

The density $f(t_{12})$ is plotted in figure 2*c* and *f* for two sets of parameter values. The density is discontinuous at τ_1 and would be discontinuous at τ_0 as well if $\theta_4 \neq \theta_5$. Note that the areas under the density over the three segments $(0, \tau_1)$, (τ_1, τ_0), and (τ_0, ∞) are the probabilities $P(G_{1a})$, $P(G_{1b})$, and $P(G_{1c}) + P(G_2) + P(G_3)$, respectively. The densities for $M \leq 0.01$ are nearly identical to that for $M = 0$, while the densities for $M = 10$ (not shown) or 1,000 are nearly the same as for $M = \infty$. The densities for $M = 0.1$ and 1 are in between. The pattern is similar to that found by Zhang et al. (2011), who examined the impact of M on species delimitation.

Note that if $M = \infty$, we have $f(t_{12}) = 1/\theta$ at $t_{12} = 0$. If M is large but finite (say, $M = 10^3$), $f(t_{12})$ has a peak near 0. From equation (3), $f(t_{12}) = 0$ at $t_{12} = 0$ if $0 \leq M < \infty$. This is because for $t_{12} = 0$ there have to be a migration and a coalescent in the small time interval $(0, \Delta t)$, an event that has probability of order $(\Delta t)^2$ if M is finite. Nevertheless, if M is large but finite, the density rises very quickly from 0 to $\sim 1/\theta$ when t_{12} increases from 0.

Results

Analysis of Simulated Data

The Null Distribution

In standard theory, the LRT statistic $2\Delta\ell$ has the asymptotic χ^2 distribution, with the degree of freedom to be the difference in the number of parameters between the two tested hypotheses. However, this large-sample theory relies on certain regularity conditions. In our LRT, those conditions are not satisfied, and the χ^2_2 is not expected to be the correct null distribution. First, the null hypothesis of our test corresponds to fixing parameter M at 0 in the alternative hypothesis, but the value 0 is at the boundary of the parameter space. Second, when $M = 0$, parameter θ becomes unidentifiable as it does not affect the likelihood. The distribution of $2\Delta\ell$ under such irregular conditions is in general unknown. Mathematical analysis of simple cases where the likelihood is given by binomial or normal distributions suggests that $2\Delta\ell$ may converge to $+\infty$ (in other words, diverge) when the data size $n \to \infty$ (Hartigan 1985). However, the rate of convergence is very slow, at log log n (Liu and Shao 2004), so that the asymptotic theory has virtually no relevance to analysis of practical data sets. Computer simulations in general confirm that use of the simple χ^2 makes the test conservative (Chen and Chen 2001).

We have conducted a small simulation under M0 to examine the null distribution of our LRT. We used four sets of parameter values. The first two sets are as in Yang (2010), based roughly on estimates from the hominoids (Burgess and Yang 2008) and the mangroves (Zhou et al. 2007). Sets 3 and 4 have larger parameter values and also different values for the three θs. The JC69 mutation model, with constant rate among loci, is used both to simulate and to analyze the data. Given the parameter values, the probabilities of the five site patterns are calculated using equation (10), and the counts of sites at each locus (n_{i0}, n_{i1}, n_{i2}, n_{i3}, n_{i4}) are generated by sampling from the multinomial distribution. Each replicate data set consists of L loci, of 500 bp each, and is analyzed under models M0 and SIM3s to calculate the test statistic $2\Delta\ell = 2(\ell_1 - \ell_0)$.

The results are summarized in table 1. We focus on two features of the null distribution. First, we examine the false positive rate when $\chi^2_{2,5\%} = 5.99$ is used to conduct the test. For all parameter values and data sizes examined here, use of the χ^2_2 makes the test very conservative, with the false positive rate well below the nominal 5% (table 1). At 15,000 loci, the MLEs of parameters θ_4, θ_5, τ_0, τ_1 under M0 are very close to the true values, but the false positive rate of the test is still far away from 5% and is instead close to 0.5–1%. There is no doubt that the χ^2_2 is not the correct null distribution. Second, we are interested in whether $2\Delta\ell$ converges to the same distribution, independent of the values of parameters in the null hypothesis, as is the case with the χ^2 under regular conditions. As the number of replicates we used (1,000) is too small to estimate a full distribution, we calculated the proportion of replicates in which $2\Delta\ell = 0$ and also the 95% percentile. Both vary depending on the data size and on the parameters in the null model (table 1). We conclude that it is not feasible to use a fixed significance value to conduct the test independent of the parameter values. It should be pointed out that one can use parametric bootstrap to estimate the correct null distribution, simulating data sets using parameter estimates under the null hypothesis. This approach is expensive and not used in this study. Instead, we use the simple χ^2_2. We note that the problem of uncertain null distribution will

Table 2. False Positive Rate of the LRT in Presence of Recombination.

Recombination Rate	$L = 10$	100	1,000	15,000
$\rho = 0.148 \times 10^{-3}$	0.0%	0.0%	0.5%	1.5%
$\rho = 0.3 \times 10^{-3}$	0.0%	0.0%	0.5%	2.0%
$\rho = 2 \times 10^{-3}$	0.0%	0.0%	2.0%	4.5%
$\rho = 8 \times 10^{-3}$	0.0%	0.1%	4.7%	27.1%

NOTE.—Recombination rate $\rho = 4Nr$ is per generation per base pair. Data are simulated using MS and SEQ-GEN under model M0 using the hominoid set of parameters: $\theta_4 = \theta_5 = 0.005$, $\tau_0 = 0.006$, and $\tau_1 = 0.004$, with recombination rate ρ, and are analyzed using the test of this paper ignoring recombination. Sequence length is 500 bp at each locus. The number of replicates is 1,000.

disappear when we extend the method to include loci with two or more sequences from either species 1 or 2 (see below).

The Impact of Recombination

Our model assumes free recombination between loci and no recombination among sites within the same locus. To examine whether recombination between sites within a locus may cause excessive false positives, we simulated data using the programs MS (Hudson 2002) to generate the genealogical trees for different sequence segments at each locus and SEQ-GEN to generate sequence alignments. The data were then analyzed using the 3s program. No reliable estimates of recombination rate are available for the mangroves, so we focussed on the parameter set for the hominoids. The recombination rate used is based on the estimates for humans, at $r = 0.37$ cM/Mb (with the 95% CI to be 0.27–0.47) (Arnheim et al. 2007). With the effective population size of $N = 10^4$, this gives $\rho = 4Nr = 0.148$ per generation per kilo base pair or 0.074 for a 500 bp locus. We include three larger values, so that we used $\rho = 0.148$, 0.3, 2, and 8 per kilo base pair. Recombination appears to be concentrated at hot spots (Myers et al. 2005; Wang and Rannala 2009). However, MS does not simulate under a model of background recombination with hotspots, as it assumes a constant recombination rate. For example, the following commands generate a replicate data set of 1,000 loci:

ms 3 1000 -T -r 0.074 500 -I 3 1 1 1 -ej 0.8 2 1 -ej 1.2 3 1|tail -n +4| grep -v // > treefile

seq-gen -mHKY -l 500 -s 0.005 -p 100 < treefile > seqfile.

The results are shown in table 2. The false positive rate of the LRT depends on the number of loci (L) and the recombination rate (ρ). For values of ρ realistic for humans ($\rho < 1$), the false positive rate is well below 5%. However, with $L = 15{,}000$ loci and $\rho = 8$ per kilo base pair, the false positive rate is as high as 27%. Overall, our test appears to be less sensitive to recombination than a similar test of gene flow using data from two species developed by Yamamichi et al. (2012). Those authors used $\theta = 0.00405$ and $\rho = 2$, 4, and 8 per kilo base pair in their simulations and found that their test was very sensitive to recombination and to mutation rate variation among loci. The differences appear to be due to several factors. First, our test uses three instead of two sequences, so that the gene tree topology and branch lengths provide information and the test is more robust. Second, our use of χ^2_2 makes our test very conservative. Nevertheless, the null model in our test is violated in presence of recombination, and the false positive rate may be high if ρ is large and if a huge number of loci are analyzed.

Table 3. Power of the LRT.

Parameter Values	$L = 10$	100	1,000	15,000
Hominoid set	0.1%	4.1%	68.8%	94.9%
Mangroves set	1.7%	31.9%	93.9%	98.5%

NOTE.—Data are simulated under model M1 (SIM3s) using two sets of parameters: $\theta_4 = \theta_5 = \theta = 0.005$, $\tau_0 = 0.006$, $\tau_1 = 0.004$ (hominoid) and $\theta_4 = \theta_5 = \theta = 0.01$, $\tau_0 = 0.02$, $\tau_1 = 0.01$ (mangroves), with the migration rate $M = 1$. The LRT is conducted using the χ^2_2 distribution at the 5% level (critical value at 5.99). Sequence length is 500 bp at each locus. The number of replicates is 1,000.

Power of the LRT

We then generated data sets under the SIM3s model with migration to examine the power of the test. We use $\theta = \theta_4 = \theta_5$ (that is, 0.005 for the hominoid set and 0.01 for the mangroves set), and $M = 1$, with on average one migrant individual per generation. The results are shown in table 3. With 10 loci, the test has virtually no power. Even with 100 loci, the power is low (4.1% for the hominoid set and 31.9% for the mangroves set). The power is fairly high when 1,000 loci are included in the data set. The large differences between the two sets of parameters are mainly due to the near 2-fold difference in mutation rate and thus the information content in the sequence data.

The power of the test based on the new SIM3s model appears to be even lower than the approximate test based on a beta distribution of τ_1 (Yang 2010: table 2). Several factors may be responsible for the differences. First, there may not be a simple correspondence between parameter q in the beta model used by Yang (2010) and the migration rate M used in this study. Second, the beta model may have captured the main differences between the null model of no gene flow and the alternative model of gene flow. Third, as discussed early, the use of χ^2_2 for the test, which costs 2 degrees of freedom, has made the new test very conservative, while the approximate test used <1 df. A useful way of improving the power of the test is to include some loci with two or three sequences from each species. Inclusion of such data will also remove the problem of nonidentifiability (see below).

Analysis of the Hominoid Data Set

Here, we apply the new test based on the SIM3s model to the genomic sequences of the human, chimpanzee, and gorilla compiled and analyzed by Burgess and Yang (2008; see also Yang 2010). These data are an update of the data of Patterson et al. (2006), with more stringent filters to remove the error-prone ends of whole-genome shotgun reads, as well as coding regions, repeats, RNA genes, and low-complexity regions. Like the model of this paper, the multispecies coalescent model (Rannala and Yang 2003) used by Burgess and Yang (2008) assumed free recombination between loci and no recombination between sites

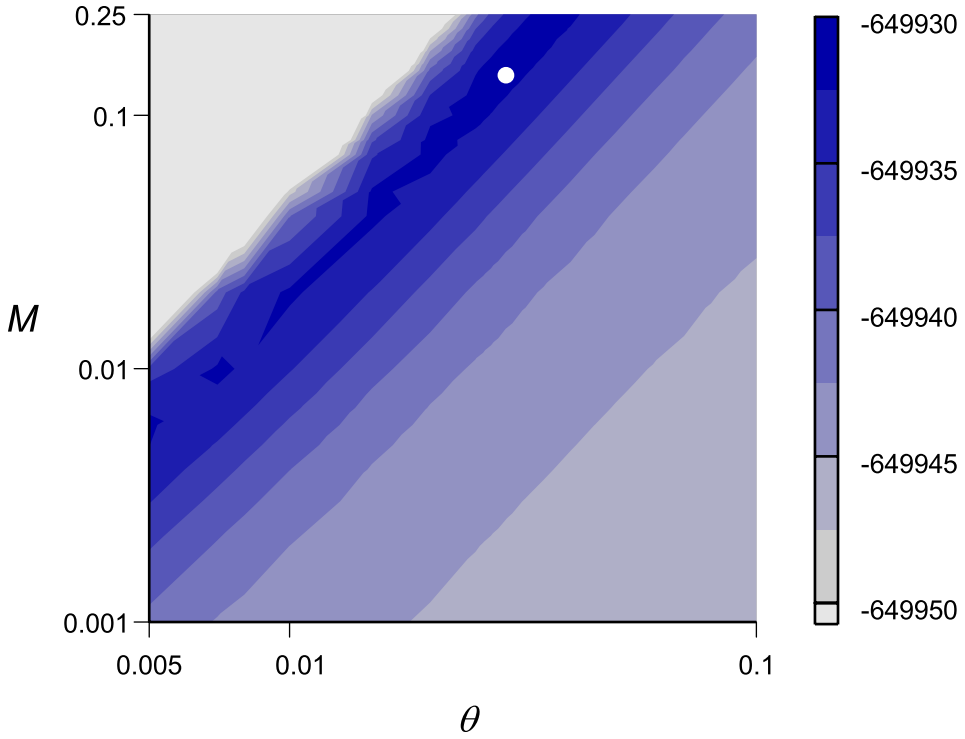

FIG. 3. A contour plot for parameters θ and M, with the other four parameters (θ_4, θ_5, τ_0, τ_1) fixed at their MLEs. The 9861 autosomal loci from the human, chimpanzee, and gorilla are analyzed under the SIM3s model. The MLEs are $\hat{\theta}_4$=0.00362, $\hat{\theta}_5$=0.00369, $\hat{\tau}_0$=0.00659, $\hat{\tau}_1$=0.00459, $\hat{\theta}$=0.0260, $\hat{M}$=0.1251, indicated by the white dot in the plot, with $\ell = -649931.90$.

in each locus. Thus Burgess and Yang (2008) used short loci with about 500 bp (so that recombination within locus is rare) and at least 10 kb of separation between loci (so that the loci are nearly freely recombining). We use the 9,861 autosomal loci and 510 X-linked loci for human (H), chimpanzee (C), and gorilla (G), as analyzed in table 3 of Yang (2010). Sites with alignment gaps and ambiguity nucleotides are removed. The median sequence length at each locus is 508 bp. As there is little evidence for mutation rate variation among those presumably neutral loci and accommodating rate variation among loci was found to have little effect in the previous analysis (Burgess and Yang 2008; Yang 2010: table 3), we assume here the same mutation rate for all loci.

When all loci on each human chromosome are analyzed as one data set, the test was not significant for any of the 22 autosomal chromosomes or the X chromosome. This is in contrast to the analysis using the approximate test based on the beta model, which is significant at 3 of the 22 autosomes (Yang 2010). This may reflect the different power of the two tests. When all the 9,861 autosomal loci are analyzed as one data set, the LRT statistic is $2\Delta\ell = 2(\ell_1 - \ell_0) = 9.43$, so that the test is significant, indicating gene flow, consistent with Yang (2010). The MLEs under the SIM3s model are listed in table 5 (column headed "Full Data"). The approximate standard errors are calculated by numerical computation of the Hessian matrix at the MLEs. These are all very small: in analysis of such genomic data sets, sampling errors are small, and systematic errors due to model violations are much more important. Figure 3 shows the log likelihood surface as a function of parameters θ and M, with the other four parameters (θ_4, θ_5, τ_0, τ_1) fixed at their MLEs. There is strong positive

Table 4. The Correlation Coefficients between MLEs of Parameters under the SIM3s Model.

	θ_{HCG}	θ_{HC}	τ_{HCG}	τ_{HC}	θ	M
θ_{HCG}						
θ_{HC}	−0.41					
τ_{HCG}	−0.73	0.37				
τ_{HC}	0.32	−0.93	−0.22			
θ	0.05	−0.18	−0.08	0.05		
M	0.25	−0.51	−0.22	0.48	0.24	

NOTE.—The 9,861 hominoid autosomal loci are analyzed. High correlations are underlined.

Table 5. Parameter Estimates Obtained from the Data Set of 9,861 Autosomal Loci from Hominoids under the SIM3s Model.

MLEs	Full Data	Left Half	Right Half
$\hat{\theta}_4$	0.00362 ± 0.00009	0.00432 ± 0.00012	0.00420 ± 0.00013
$\hat{\theta}_5$	0.00369 ± 0.00026	0.00245 ± 0.00043	0.00319 ± 0.00056
$\hat{\tau}_0$	0.00659 ± 0.00004	0.00625 ± 0.00006	0.00628 ± 0.00007
$\hat{\tau}_1$	0.00459 ± 0.00008	0.00493 ± 0.00018	0.00464 ± 0.00020
$\hat{\theta}$	0.02601 ± 0.00140	0.02114 ± 0.00152	0.02851 ± 0.00291
$\hat{M}$	0.1250 ± 0.0065	0.1250 ± 0.0491	0.1250 ± 0.0558
ℓ_1	−649,931.90	−325,045.125	−325,035.85
$2\Delta\ell$	9.43	3.25	0.97

NOTE.—Left Half refers to the data set in which only the first half of the sites at each locus are used, whereas Right Half refers to the data set in which only the second half of the sites at each locus are used.

correlation between $\hat{\theta}$ and $\hat{M}$, especially at the logarithmic scale. Table 4 shows the correlations between the MLEs of parameters.

Burgess and Yang (2008: table 6) examined the impact of recombination within locus by analyzing reduced data sets in which each locus was shortened. Here, we generated two reduced data sets, consisting of either the left half or the right half of every locus in the data set of 9,861 autosomal loci. The parameter estimates and LRT statistics for those two reduced data sets are listed in table 5. The test is not significant for either of the reduced data set. This can be due to two reasons. The first is that each of the two reduced data sets has half as much data and contains less information than the full data set. The second is that recombination may be frequent and cause a false positive in the full data set, while it is unimportant in the reduced data set. Although both factors may have contributed to the difference, the simulation results discussed above demonstrate that the false positive rate for the test is low even when realistic amounts of recombination. It thus appears likely that the significant result in the full data sets is not a false positive. Parameter estimates (table 5) are similar across the three data sets. It is interesting to note that $\hat{M}=0.125$ in all three analyses.

Discussion

A common form of model nonidentifiability is overparametrization, or the use of more parameters than can be estimated from the data. For example, in the problem of distance estimation between two sequences, the likelihood is a function of the sequence distance $d = t \cdot r$ but not of the time t and rate r separately. We then use d as the single parameter in the model, instead of two parameters t and r. The nonidentifiability of the SIM3s model is different in that the model is not overparametrized. If we use a and b as parameters, the model will become identifiable, although a and b do not have the simple biological interpretations that parameters θ and M have.

The model is not identifiable because the data contain one sequence from each of the three species. However, if some loci with other sample configurations are included, for example, loci with two or three sequences from the same species, the model will become identifiable. We are now working to extend the SIM3s model to deal with loci of arbitrary configurations. Such loci may be represented by different initial states in the Markov chain for epoch E_1, such as 111, 222, 333, 112, 113, 122, 133, 223, 233, 11, 22, 33, 12, 13, and 23. The extension will have several benefits. First, the data will be much more informative about θ_1 and θ_2, and will in turn help the estimation of the migration rates. The model will then become identifiable. Second, as parameters θ_1 and θ_2 are identifiable in both hypotheses, the null distribution will have a known well-behaved distribution, that is, a 1:1 mixture between 0 and χ^2_1. Third, inclusion of such loci will make the LRT more powerful. Fourth, such data will enable implementation of more complex models with arbitrary patterns of migration.

To extend the SIM3s model to consider such loci, one has to deal with much larger rate matrices $Q^{(1)}$ and $Q^{(2)}$. The integrals involved in the likelihood function will be 2-D if three sequences are available at the locus, or 1-D for two sequences, so that the computation should be feasible. Dealing with data of four or more sequences at a locus will pose much greater difficulties. First, the state space of the Markov chain increases quickly and so is the computation involved in calculating the transition probability matrix $P(t)$. Second, the summation over the gene tree topologies and the integration over the coalescent times will be much more expensive, as there will be many more possible gene trees to sum over, and as the dimension of the integrals will become 3 or greater.

Another extension of the model is to deal with nonhomogeneous migration rates over time. The rate of gene flow may be expected to be high at the initial stage of speciation and to decrease when the two populations diverge to become separate species. For example, one may assume a constant migration rate M since species divergence until a time point T, $0 < T < \tau_1$, when gene flow ceases. Both the migration rate M and the time point T will be parameters in the model. The distribution of gene trees and coalescent times under this model can be derived in a straightforward manner using the Markov chain characterization of this study by breaking time epoch E_1 into two segments: E_{1a}: $0 < t < T$ and E_{1b}: $T < t < \tau_1$. Another model may assume that the migration rate has been decreasing at an exponential rate β since species divergence, so that the migration rate at time t is $M_0 \cdot \exp\{\beta(\tau_1 - t)\}$, for $0 < t < \tau_1$. The initial migration rate M_0 and the exponential rate β will be parameters in the model, to be estimated from the data. It appears that this model can be implemented in a similar way to the treatment of deterministically varying population size in the coalescent model (Slatkin and Hudson 1991; Griffiths and Tavaré 1994). Those two models may be more realistic than the constant-migration model of this study in describing the gradual buildup of reproductive isolation after species separation.

Acknowledgments

We thank Hilde Wilkinson-Herbots, Peter Beerli, and an anonymous referee for many constructive comments and Hideki Inan for discussions of the impact of recombination on the test. This study is supported by a grant from the Biotechnological and Biological Sciences Research Council (BBSRC BB/G006431/1). Z.Y. is a Royal Society-Wolfson Merit Award holder.

Appendix

Here, we derive the transition probability matrix $P(t)$ from the Q matrix for time epoch 1 (eq. 1). The spectral decomposition of Q is derived to be

$$Q = U\Lambda U^{-1}, \tag{24}$$

where $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$,

$$U = \begin{bmatrix} 1 & -1 & -1 & 1 & 1 \\ 1 & 0 & 0 & \frac{c-a}{4w} & \frac{c+a}{4w} \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix}, \quad (25)$$

and

$$U^{-1} = \begin{bmatrix} 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & 0 & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{4}\left(1+\frac{c}{a}\right) & -\frac{2w}{a} & \frac{1}{4}\left(1+\frac{c}{a}\right) & -\frac{c+a-4w}{4a} & -\frac{c+a-4w}{4a} \\ \frac{1}{4}\left(1-\frac{c}{a}\right) & \frac{2w}{a} & \frac{1}{4}\left(1-\frac{c}{a}\right) & \frac{c-a-4w}{4a} & \frac{c-a-4w}{4a} \end{bmatrix} \quad (26)$$

with $a = \sqrt{c^2 + 16w^2}$.

Thus the transition probability matrix is given by

$$P(t) = U \cdot e^{\Lambda t} \cdot U^{-1}, \quad (27)$$

where $e^{\Lambda t} = \mathrm{diag}\{e^{\lambda_1 t}, e^{\lambda_2 t}, e^{\lambda_3 t}, e^{\lambda_4 t}, e^{\lambda_5 t}\}$.

References

Arnheim N, Calabrese P, Tiemann-Boege I. 2007. Mammalian meiotic recombination hot spots. *Annu Rev Genet.* 41:369–399.

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

Barton NH. 2006. Evolutionary biology: how did the human species form? *Curr Biol.* 16:R647–R650.

Beerli P. 2006. Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics* 22:341–345.

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

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

Burgess R, Yang Z. 2008. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Mol Biol Evol.* 25:1979–1994.

Chen H, Chen J. 2001. The likelihood ratio test for homogeneity in finite mixture models. *Can J Stat.* 29:201–215.

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

Hartigan JA. 1985. A failure of likelihood asymptotics for normal mixtures. Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II; 1983; Berkeley, CA. Belmont (CA): Wadsworth Statistics/Probability Series, Wadsworth. p. 807–810.

Hey J. 2010a. Isolation with migration models for more than two populations. *Mol Biol Evol.* 27:905–920.

Hey J. 2010b. The divergence of chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses. *Mol Biol Evol.* 27:921–933.

Hey J, Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* 167:747–760.

Hobolth A, Andersen LN, Mailund T. 2011. On computing the coalescence time density in an isolation-with-migration model with few samples. *Genetics* 187:1241–1243.

Hudson RR. 1983. Testing the constant-rate neutral allele model with protein sequence data. *Evolution* 37:203–217.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press. p. 21–123.

Liu X, Shao Y. 2004. Asymptotics for the likelihood ratio test in a two-component normal mixture model. *J Stat Plan Inference.* 123:61–81.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310:321–324.

Nath HB, Griffiths RC. 1993. The coalescent in two colonies with symmetric migration. *J Math Biol.* 31:841–852.

Notohara M. 1990. The coalescent and the genealogical process in geographically structured populations. *J Math Biol.* 29:59–75.

Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* 441:1103–1108.

Rannala B. 2002. Identifiability of parameters in MCMC Bayesian inference of phylogeny. *Syst Biol.* 51:754–760.

Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164:1645–1656.

Slatkin M, Hudson RR. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129:555–562.

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

Wang Y, Rannala B. 2009. Population genomic inference of recombination rates and hotspots. *Proc Natl Acad Sci U S A.* 106:6215–6219.

Wang Y, Hey J. 2010. Estimating divergence parameters with small samples from a large number of loci. *Genetics* 184:363–379.

Wilkinson-Herbots HM. 1998. Genealogy and subpopulation differentiation under various models of population structure. *J Math Biol.* 37:535–585.

Wilkinson-Herbots HM. 2008. The distribution of the coalescence time and the number of pairwise nucleotide differences in the "isolation with migration" model. *Theor Popul Biol.* 73:277–288.

Yamamichi M, Gojobori J, Innan H. 2012. An autosomal analysis gives no genetic evidence for complex speciation of humans and chimpanzees. *Mol Biol Evol.* 29:145–156.

Yang Z. 1994. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Syst Biol.* 43:329–342.

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

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

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.

Yang Z. 2010. A likelihood ratio test of speciation with gene flow using genomic sequence data. *Genome Biol Evol.* 2:200–211.

Zhang C, Zhang D-X, Zhu T, Yang Z. 2011. Evaluation of a Bayesian coalescent method of species delimitation. *Syst Biol.* 60:747–761.

Zhou R, Zeng K, Wu W, Chen X, Yang Z, Shi S, Wu C-I. 2007. Population genetics of speciation in nonmodel organisms: I. ancestral polymorphism in mangroves. *Mol Biol Evol.* 24:2746–2754.