27 May 2016

Maximum Likelihood Implementation of an Isolation-with-Migration Model for Three Species

DANIEL DALQUEN[1*], TIANQI ZHU[2*] and ZIHENG YANG[1, 2, 3]

[1] *Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK*

[2] *Center for Computational Genomics, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*

Running head: IMPLEMENTATION OF IM MODEL FOR 3 SPECIES

[*] Those authors contributed equally to the study.

Key words: Multispecies coalescent, maximum likelihood, speciation, IM model, migration.

[3]*Correspondence to*: Ziheng Yang
Address: Department of Genetics, Evolution and Environment
University College London
Darwin Building
Gower Street
London WC1E 6BT
England
Email: z.yang@ucl.ac.uk
Phone: +44 (20) 7679 4379

Abstract.— We develop a maximum likelihood (ML) method for estimating migration rates between species using genomic sequence data. A species tree is used to accommodate the phylogenetic relationships among three species, allowing for migration between the two sister species, while the third species is used as an outgroup. A Markov chain characterization of the genealogical process of coalescence and migration is used to integrate out the migration histories at each locus analytically, while Gaussian quadrature is used to integrate over the coalescent times on each genealogical tree numerically. This is an extension of our early implementation of the symmetrical isolation-with-migration model for three species to accommodate arbitrary loci with two or three sequences per locus and to allow asymmetrical migration rates. Our implementation can accommodate tens of thousands of loci, making it feasible to analyze genome-scale datasets to test for gene flow. We calculate the posterior probabilities of gene trees at individual loci to identify genomic regions that are likely to have been transferred between species due to gene flow. We conduct a simulation study to examine the statistical properties of the likelihood ratio test for gene flow between the two ingroup species and of the maximum likelihood estimates of model parameters such as the migration rate. Inclusion of data from a third outgroup species is found to increase dramatically the power of the test and the precision of parameter estimation. We compiled and analyzed several genomic datasets from the *Drosophila* fruit flies. Our analyses suggest no migration from *D. melanogaster* to *D. simulans*, and a significant amount of gene flow from *D. simulans* to *D. melanogaster*, at the rate of ~0.02 migrant individuals per generation. We discuss the utility of the multispecies coalescent model for species tree estimation, accounting for incomplete lineage sorting and migration.

Migration or gene flow is an important biological process that affects our interpretation of genetic data from both within and between species (e.g., Patterson et al., 2006; Innan and Watanabe, 2006; Yamamichi et al., 2012; Leaché et al., 2013; Mallet et al., 2016). For example, different models of speciation make different predictions about the presence or absence of gene flow at the time of species formation. There is a rich body of literature in population genetics concerning models of population subdivision and migration, starting from Wright (1931; 1943). For example, in the finite-island model, any population can exchange migrants with any other (Wright, 1943), while in the stepping-stone model, only neighboring populations can exchange migrants (Kimura and Weiss, 1964). The standard single-population coalescent theory (Kingman, 1982) has been extended to deal with such models of population structure and migration, in the so-called *structured coalescent* (e.g., Li, 1976; Strobeck, 1987; Takahata, 1988; Notohara, 1990; Nath and Griffiths, 1993; Wilkinson-Herbots, 1998). Models of population structure have been implemented in computer programs such as GENETREE (Bahlo and Griffiths, 2000) and MIGRATE (Beerli and Felsenstein, 1999; 2001; Beerli, 2006), which allow joint estimation of population sizes and migration rates from genetic data.

However, population structure models ignore the phylogenetic relationships among the populations and their divergence times. The isolation-with-migration (IM) model is attractive as it incorporates the population/species phylogeny in a model of migration. They allow us to estimate the migration rates and other parameters such as the species divergence times and population sizes under more realistic models (Nielsen and Wakeley, 2001; Hey and Nielsen, 2004; Wilkinson-Herbots, 2008; 2012). Another yet unexplored use of the IM model is species tree estimation under the multispecies coalescent model with migration, accounting for both incomplete lineage sorting and introgression. Coalescent-based phylogenetic inference, which accommodate gene tree-species tree discordance due to incomplete lineage sorting, has been heralded as a paradigm shift in molecular phylogenetics (Edwards, 2009). Recent analyses of genomic datasets have found widespread conflicts among nuclear gene trees and between the mitochondrial gene tree and the nuclear species tree, for example, in mosquitos (Fontaine et al., 2015), butterflies (Martin et al., 2013), frogs (Zhou et al., 2012), birds (Ellegren et al., 2012), hares (Melo-Ferreira et al., 2012), bears (Liu et al., 2014; Kutschera et al., 2014), and gibbons (Chan et al., 2013). Hybridization both between sister species and between non-sister species is commonly observed between modern species, so it is natural to expect it to have occurred in ancestral species as well, especially during adaptive radiations (Mallet, 2005; Mallet et al., 2016). Many empirical studies have highlighted incomplete lineage sorting (or rapid radiation) and gene flow (introgression) as the two major challenges to species tree estimation when the species are closely related. While the multispecies coalescent model with gene flow should accommodate both factors naturally, full likelihood methods of species tree estimation under the model are currently lacking.

Full likelihood implementation of the IM model for the analysis of genetic sequence data is challenging because calculation of the likelihood function has to average over the genealogical history at every locus, which includes the gene tree topology, the branch lengths (the coalescent times), and the whole migration trajectory (the number, directions and times of all migration events). The IM programs (Nielsen and Wakeley, 2001; Hey and Nielsen, 2004; Hey, 2010), for example, are not practical for analyzing datasets with a few hundred loci (Hey, 2010). Approximations are often necessary to analyze genome-scale data with many loci (Gronau et al., 2011).

When there are only a few sequences at a locus, it is possible to integrate out the migration history either numerically or analytically (Wang and Hey, 2010; Lohse et al., 2011; Zhu and Yang, 2012; Andersen et al., 2014). It is then feasible to analyze tens of thousands of loci even though only a few sequences are sampled at each locus. Here loci may be defined as loosely linked short genomic segments that are far apart from each other, so that recombination within a locus is unlikely to affect the gene tree distribution, while different loci are nearly independent due to recombination events (Burgess and Yang, 2008; Lohse et al., 2011). Wang and Hey (2010) used numerical integration and special functions to integrate out the migration history under the IM model for two species when the data at every locus consist of two sequences, with one from each species. A more efficient approach is to integrate out the migration trajectory analytically by using the Markov chain characterization of the coalescent process with migration developed in the structured coalescent framework (Notohara, 1990; Nath and Griffiths, 1993; Hobolth et al., 2011; Zhu and Yang, 2012; Andersen et al., 2014). For example, with only two sequences at a locus, the probability of the sequence data at any locus depends on the sequence divergence time *t* only, and not on the number and times of the migration events. The density for *t* can be calculated analytically (Hobolth et al., 2011; see also Nath and Griffiths, 1993; Wilkinson-Herbots, 2008). Lohse *et al.* (2011) derived probabilistic distributions of gene trees using generating functions and symbolic algebra in Mathematica. The implementation allows more than two sequences at each locus, thus increasing the power of the analysis (Lohse et al., 2011).

Zhu and Yang (2012) implemented the IM model for three species, assuming symmetry in the migration rates and population sizes between species 1 and 2 (with $M_{12} = M_{21} = M$, and $\theta_1 = \theta_2$), while a third species (species 3) is used as the outgroup. They constructed a likelihood ratio test (LRT) by comparing this model, M2 (gene flow), with a null model of no migration with $M = 0$ (M0: no gene flow). In their implementation, the data at every locus are assumed to consist of three sequences, with one sequence from each species (this data configuration is referred to in this paper as '123'). This restriction on data leads to reduced power of the test and to an unusual case of unidentifiability (Zhu and Yang, 2012). Recently, Andersen *et al.* (2014) have considered the IM model in a general setting, in which one ancestral species splits into an arbitrary number of populations at a time in the past (so that the populations are related by a star phylogeny), allowing for migration between any two populations. The authors developed a strategy for 'lumping' states in the Markov chain to alleviate the problem of state-space explosion. Their implementation, for the case of two diploid individuals from two species (four sequences per locus), assumed free recombination between any two sites (alignment columns). Under this assumption, the data at different sites are independent (conditional on the species phylogeny and parameters in the model) so that the sequence dataset can be summarized as counts of 4^4 possible site patterns (nucleotide combinations), and the authors were able to integrate out the coalescent times in the gene trees for each site analytically (Andersen et al., 2014, sections 5 and 8.4).

In this study we extend the implementation of Zhu and Yang (2012). Like many previous studies such as Takahata et al. (1995), Wang and Hey (2010), and Lohse et al. (2011), we work under the assumption of complete linkage within a locus and free recombination between loci. We note that both free recombination and complete linkage within a locus are extreme assumptions, and their impact on the inference is not yet well understood (but see Burgess and Yang, 2008; Zhu and Yang, 2012). We accommodate loci of two or three sequences of arbitrary configurations, including '11' (two sequences from species 1), '112' (two sequences from species 1 and one sequence from species 2), and so on. Extension to arbitrary loci (with two or three sequences per locus) improves the power of the likelihood ratio test of gene flow and makes it possible to estimate the migration rates, which are unidentifiable with '123' loci alone (Zhu and Yang, 2012). We focus on migration between species 1 and 2, and include species 3 as an outgroup to improve the power of the analysis. As nicely discussed by Lohse *et al.* (2011), the outgroup may be informative about the gene tree topology as well as the branch lengths and about the ancestral nucleotide states in the common ancestor of species 1 and 2. Inclusion of the outgroup may also make the inference more robust to mutation rate variation among loci (Yang, 2002). We remove the symmetry assumption of the model, so that the inference can be conducted under a more realistic model. We develop an empirical Bayes approach to calculating the posterior probabilities of gene tree topologies at individual loci, which may be informative about whether the locus has been transferred between species due to gene flow. We conduct a simulation study to examine the false positive rate and power of the LRT of gene flow as well as the bias and variance of maximum likelihood estimates of model parameters. We use the genome sequences of *Drosophila melanogaster, D. simulans*, and *D. yakuba* to construct multi-locus datasets and apply our new method to infer the pattern and rate of migration between those fruit-fly species.

THEORY AND METHODS

Model and Data

The terms species and population are used interchangeably in this paper. The species tree is ((1, 2), 3), with 4 and 5 to be the ancestral species (Fig. 1a). The two divergence events on the species tree define three time epochs: E_1: (0, τ_1), E_2: (τ_1, τ_0) and E_3: (τ_0, ∞) (Fig. 1a). We consider two models. M0 (no gene flow) assumes no gene flow and is the multispecies coalescent model for three species (Takahata et al., 1995; Yang, 2002; Rannala and Yang, 2003). Model M2 (gene flow) allows migration between species 1 and 2 (during time epoch E_1), but not from or to species 3.

There are nine parameters in the general IM model for three species, including two species divergence times (τ_0 and τ_1), five effective population sizes (θ_1, θ_2, θ_3, θ_4, θ_5), and two migration rates (M_{12} and M_{21}). Here τ_0 and τ_1 are scaled by the mutation rate and are measured by the expected number of mutations per site, and $\theta_i = 4N_i\mu$ (i = 1, …, 5) are the population size parameters for the five species, with N_i being the (effective) population size of species i and μ the mutation rate per site per generation. The migration rate is $M_{ij} = N_jm_{ij}$, where m_{ij} is the proportion of individuals in population j that are immigrants from population i. We define parameters by referring to the real-world process with time running forward (rather than the coalescent view with time running backward) so that M_{ij} is the expected number of migrant individuals from populations i to j per generation. The parameters under M2 (gene flow) are $\Theta_2 = \{\tau_0, \tau_1, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, M_{12}, M_{21}\}$. Model 0 (no gene flow) is a special case of M2 with $M_{12} = M_{21}= 0$, with parameters $\Theta_0 = \{\tau_0, \tau_1, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5\}$. Note that the symmetrical versions of M0 and M2 assume $\theta_1 = \theta_2$ and $M_{12} = M_{21}$ (Zhu and Yang, 2012).

The data consist of multiple neutral loci. At each locus, two or three sequences are sampled, each from any of the three species. We focus mainly on the case of three sequences at a locus. The case of two sequences is much simpler and will be described briefly. Let the three sequences at a locus be *a*, *b*, and *c*. Each sequence will also be labelled by the population it is sampled from. For example, the initial state for a locus with data configuration '123' (with one sequence from each of the three species) is recorded as $1_a2_b3_c$. The Markov chain runs backwards in time, describing the change of states due to coalescent and migration. For example a locus with initial state $1_a2_b3_c$ may enter the state $2_{ab}3_c$, which means that sequences *a* and *b* have coalesced so that only two sequences remain in the sample and the ancestor of sequences *a* and *b* is in population 2 while sequence *c* is in population 3. There are six gene tree shapes for three sequences: G_1-G_6 (Fig. 1b-g), depending on the time epochs during which the two coalescent events occur. When we keep track of both the sequence IDs (*a*, *b*, *c*) and the population IDs (1, 2, 3), each gene tree shape may correspond to three distinct gene trees (Fig. 2). For example, tree shape G_6 corresponds to three gene trees: G_{6c}: ((*a*, *b*), *c*); G_{6a}: ((*b*, *c*), *a*); and G_{6b}: ((*c*, *a*), *b*), where the subscript is the more distantly related sequence in the gene tree. However, depending on the initial data configuration, some of the gene trees may not be possible (for example, for a '123' locus, only gene trees G_{3c}, G_{5c}, G_{6c}, G_{6a}, G_{6b} are possible under M2), and furthermore some of the gene trees have the same probability distribution under the model (such as G_{6c}, G_{6a}, and G_{6b}). To avoid excessive notation we make a distinction between gene tree shapes and gene trees only if there is a risk of confusion.

Likelihood Function for Three Sequences at a Locus

We assume that the sequences at each locus are already aligned, with alignment gaps and ambiguity nucleotides removed. We use the JC69 mutation model (Jukes and Cantor, 1969) to correct for multiple substitutions. The different loci are assumed to have the same mutation rate, although relative rates for the loci can be incorporated in the likelihood calculation (if available, for example, through comparison with an outgroup species, Yang, 2002). The sequence alignment at any locus *i* with three sequences can be summarized as the counts, $D_i = (n_0, n_1, n_2, n_3, n_4)$, of sites with five different site patterns: *xxx*, *xxy*, *yxx*, *xyx*, and *xyz*, where *x*, *y* and *z* are any distinct nucleotides. The probability of the data given the gene tree topology (G) and branch lengths (b_0, b_1) (Fig. 2), $P(D_i|G, b_0, b_1)$, is thus given by the multinomial distribution, with the probabilities of the five site patterns calculated efficiently under the JC69 model (Saitou, 1988; Yang, 1994). Conveniently, $P(D_i|G, b_0, b_1)$ depends on the gene tree topology and branch lengths, but not on which time epoch each coalescent event occurs in (Yang, 2002; 2010).

The probability of data at locus *i* is an average over the gene tree topologies and coalescent times

$$f(D_i \mid \Theta) = \sum\nolimits_k \int_{l_0}^{u_0} \int_{l_1}^{u_1} P(D_i \mid G_k, b_0, b_1)\, f(G_k, t_0, t_1 \mid \Theta)\, \mathrm{d}t_1 \mathrm{d}t_0 \,, \tag{1}$$

where the sum is over all possible gene trees for the locus, while the integrals are over the coalescent times t_0 and t_1, with the integral limits $t_0 \in (l_0, u_0)$ and $t_1 \in (l_1, u_1)$ given below. Note that the branch lengths b_0 and b_1 in the gene tree are simple linear functions of t_0 and t_1 (Figs. 1 and 2 and Table 1). The probability of the genealogy, $f(G_k, t_0, t_1|\Theta)$, depends on the model (M_0 or M_2) and will be described in the next section. For data configurations with three sequences, there are up to 6 × 3 = 18 gene trees to average over.

Finally, the log likelihood of the data at all L loci, $D = \{D_i\}$, is a sum over the L loci

$$\ell(\Theta; D) = \sum_{i=1}^{L} \log f(D_i \mid \Theta). \qquad (2)$$

Note that our model assumes that the n sites in the sequence at the locus share the same genealogical tree (topology and coalescent times). This contrasts with the implementation of Andersen *et al.* (2014), which assumes that the different sites have independent histories.

Implementation of Model M_0 (No Gene Flow)

We first discuss our ML implementation of model M0, which assumes no migration between any two populations. The implementation of Yang (2002) considered '123' loci only so that the model involve only four parameters: $\Theta_0 = \{\tau_0, \tau_1, \theta_4, \theta_5\}$. Here we allow arbitrary loci of two or three sequences, with up to seven parameters in the model: $\Theta_0 = \{\tau_0, \tau_1, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5\}$. Note that the population size parameter for a modern species (θ_1, θ_2, or θ_3) exists in the model only if two or more sequences are sampled from that species at least at one locus.

Consider a locus with three sequences. In general, the probability density of the gene tree has the form

$$f(G_k, t_0, t_1) = \text{rates} \times \mathrm{e}^{-T} = \tfrac{2}{\theta_i}\tfrac{2}{\theta_j}\mathrm{e}^{-T}, \qquad (3)$$

where parameters θ_i and θ_j are for the populations in which the two coalescent events occur and the exponential term e^{-T} is the probability that no coalescent event occurs in the rest of the gene tree, with T being the *total per-lineage-pair coalescent waiting time* of Yang (2014, p.336). Note that the coalescent rate for a pair of sequences in a population with population size parameter θ is $2/\theta$: for very small Δt, the probability that the pair will coalesce during the time interval $(t, t + \Delta t)$ is $\tfrac{2}{\theta}\Delta t$.

Take, for example, configuration '111', with the initial state $1_a1_b1_c$. The probability of data for the locus (Eq. 1) is an average over 6 × 3 gene trees. For example, in the case of gene tree G_{1c}: ((a, b), c), the probability density of the gene tree (with coalescent times) is

$$f(G_{1c}, t_0, t_1) = \tfrac{2}{\theta_1}\tfrac{2}{\theta_1}\mathrm{e}^{-T} = \tfrac{2}{\theta_1}\tfrac{2}{\theta_1}\mathrm{e}^{-\frac{6}{\theta_1}t_1 - \frac{2}{\theta_1}t_0}, \quad t_0 > 0,\ t_1 > 0,\ t_0 + t_1 < \tau_1, \qquad (4)$$

where $\tfrac{2}{\theta_1}$ and $\tfrac{2}{\theta_1}$ are the rates for the two coalescent events, both occurring in species 1. Because of the symmetry of the '111' locus, the density is the same for the three gene trees: G_{1c}, G_{1a}, and G_{1b}. The densities and rates for all data configurations and gene trees are summarized in Table S1. Note that some gene trees are not possible for certain configurations of loci (e.g., gene trees G_{1c}, G_{1a}, and G_{1b} for '112' loci).

To compute the integrals of equation (1) numerically, we apply a linear transform. Let $x_0 = \frac{2}{\theta_i}t_0$ and $x_1 = \frac{2}{\theta_j}t_1$ be the coalescent times measured in generations, where θs are for the populations in which the coalescent events occur. Each integral in equation (1) then becomes

$$\int_{l_0}^{u_0}\int_{l_1}^{u_1} P(D_i\,|\,G_k,b_0,b_1)\,f(G_k,t_0,t_1)\,\mathrm{d}t_1\mathrm{d}t_0 = \int_{l_0'}^{u_0'}\int_{l_1'}^{u_1'} P(D_i\,|\,G_k,b_0,b_1)\,f(G_k,x_0,x_1)\left|\tfrac{\partial(t_0,t_1)}{\partial(x_0,x_1)}\right|\mathrm{d}x_1\mathrm{d}x_0\,. \qquad (5)$$

In several cases (gene tree shapes G_1 and G_4 for initial state '111'; G_4 for '112'; and G_1, G_2 and G_4 for '333'), the integration region is a triangle (for instance, the region for G_1 is given by $t_0 > 0$, $t_1 > 0$, $t_0 + t_1 < \tau_1$; see Fig. 1). As we calculate the 2-D integral of equation (5) by calculating two 1-D integrals using Gaussian quadrature (the so-called product rule), the integral region has to be a rectangle. We thus apply a transform to achieve this. For example, in the case of G_1 for the initial state '111', we use $x_0 = \frac{2}{\theta_1}(t_0 + t_1)$, $x_1 = \frac{t_1}{t_0+t_1}$, so that $t_0 = \frac{\theta_1}{2}x_0(1-x_1)$, $t_1 = \frac{\theta_1}{2}x_0x_1$. The new limits are $0 < x_0 < \frac{2}{\theta_1}\tau_1$, $0 < x_1 < 1$, and the Jacobi of the transform is $\left|\frac{\partial(t_0,t_1)}{\partial(x_0,x_1)}\right| = \frac{\theta_1}{2}\frac{\theta_1}{2}x_0$. Then

$$\int_0^{\tau_1}\int_0^{\tau_1-t_0} P(D_i\,|\,G_{1k},b_0,b_1)\times\tfrac{2}{\theta_1}\tfrac{2}{\theta_1}\mathrm{e}^{-\frac{6}{\theta_1}t_1-\frac{2}{\theta_1}t_0}\mathrm{d}t_1\mathrm{d}t_0 = \int_0^{\frac{2}{\theta_1}\tau_1}\int_0^1 P(D_i\,|\,G_{1k},b_0,b_1)\times x_0\mathrm{e}^{-2x_0x_1-x_0}\,\mathrm{d}x_1\mathrm{d}x_0\,, \qquad (6)$$

where $b_0 = t_0$ and $b_1 = t_1$ in the integral on the left-hand side, and $b_0 = \frac{\theta_1}{2}x_0(1-x_1)$ and $b_1 = \frac{\theta_1}{2}x_0x_1$ in the integral on the right-hand side.

The transforms from (t_0, t_1) to (x_0, x_1) are summarized in Table S2. We use Gaussian quadrature to calculate the 2-D integrals of equations (5) or (6). Except where stated otherwise, we used $K = 16$ points in the quadrature. See Yang (2010) for details. It is necessary to apply scaling to avoid underflows as the probabilities of equation (1) may be too small to represent in the computer.

The case of two sequences. In the case of two sequences at a locus, the possible initial states are 11, 12, 22, 13, 23, and 33, depending on which populations the two sequences are sampled from. The simple gene tree has two branches, which have the same length t, with density $f(t|\Theta)$ (Table 1). For instance, with the initial state 11 (two sequences from species 1), $f(t|\Theta)$ is a piecewise continuous function because the population size and thus the coalescent rate may differ in the three time epochs. The sequence data at the locus are summarized as d_i differences out of n_i sites. Then the probability of observing d_i differences at n_i sites given that the two sequences separated time t ago is

$$f(d_i|t) = \left(\tfrac{3}{4} - \tfrac{3}{4}\mathrm{e}^{-8t/3}\right)^{d_i}\left(\tfrac{1}{4} + \tfrac{3}{4}\mathrm{e}^{-8t/3}\right)^{n_i-d_i}. \qquad (7)$$

The (unconditional) probability of observing the data at the locus is an average over the coalescent time

$$f(d_i|\Theta) = \int_0^\infty f(t\,|\,\Theta)\,f(d_i\,|\,t)\,\mathrm{d}t\,. \qquad (8)$$

Gaussian quadrature is used to calculate the 1-D integral, with the transform $x = \frac{2}{\theta_j}t$ (Table 1).

Implementation of Model M2 (gene flow)

Under model M2 (gene flow), the likelihood is given by equation (1) as before, and the probability of the data at each locus $P(D_i|G_k, b_0, b_1)$ remains the same. However, the probability density for the gene trees, $f(G_k, t_0, t_1)$, depends on the migration rates and differs from that under model M0. Our aim in this section is thus to describe the gene-tree density. We use a Markov chain to characterize the process of coalescent and migration when we trace the gene genealogy backwards in time. In the general case, the states of the Markov chain will include both the population IDs and sequence IDs. Because of our assumption of no migration involving species 3, the coalescent process during time epochs E_2 and E_3 are essentially the standard single-population coalescent. Thus, we focus on epoch E_1. While it is possible to use one Markov chain for all initial states, we use different Markov chains depending on the initial states to increase computational efficiency (Table *2*). The Markov chain characterization allows one to calculate the probability density for the gene tree topology and coalescent times, $f(G_k, t_0, t_1)$, with the migration history integrated out analytically (Hobolth et al., 2011; Zhu and Yang, 2012; Andersen et al., 2014). We do not use the idea of Andersen et al. (2014) for lumping states in the Markov chain because it would add much complexity to the algorithm with no or little gain for the cases of two or three sequences per locus. For the general migration case with three species, lumping actually increases the number of states from 12 to 15 for two sequences, and from 57 to 70 for three sequences (Andersen et al., 2014, table 2). We note that for four or more sequences per locus, Andersen et al.'s algorithm may lead to considerable reduction of the state space.

We illustrate the theory using gene tree G_{1c}: $((a, b), c)$ and initial state s = '111'. We take advantage of the symmetry of the initial state and consider a reduced Markov chain with eight states, dropping the sequence IDs: {111, 112, 122, 222, 11, 12, 22, 1|2} (Table 2). Here the state '1|2' means one sequence in either population 1 or 2. When both coalescent events have occurred and there is only one sequence in the sample, there will be no need to keep track of the population ID, so that states 1 and 2 can be lumped into one artificial absorbing state (Andersen et al., 2014). The rate matrix is given in Table 3. For gene tree shape G_1, we have $f(G_{1c}, t_0, t_1) = f(G_{1a}, t_0, t_1) = f(G_{1b}, t_0, t_1) = \frac{1}{3} f(G_1, t_0, t_1)$, with

$$\begin{aligned} f(G_1,t_0,t_1) = {} & 3\tfrac{2}{\theta_1} P_{s,111}(t_1)\left(\tfrac{2}{\theta_1} P_{11,11}(t_0) + \tfrac{2}{\theta_2} P_{11,22}(t_0)\right) + \tfrac{2}{\theta_1} P_{s,112}(t_1)\left(\tfrac{2}{\theta_1} P_{12,11}(t_0) + \tfrac{2}{\theta_2} P_{12,22}(t_0)\right) \\ & + \tfrac{2}{\theta_2} P_{s,122}(t_1)\left(\tfrac{2}{\theta_1} P_{12,11}(t_0) + \tfrac{2}{\theta_2} P_{12,22}(t_0)\right) + 3\tfrac{2}{\theta_2} P_{s,222}(t_1)\left(\tfrac{2}{\theta_1} P_{22,11}(t_0) + \tfrac{2}{\theta_2} P_{22,22}(t_0)\right). \end{aligned} \tag{9}$$

Note that the probability density function here has the interpretation that $f(G_1, t_0, t_1)\Delta t_0 \Delta t_1$, for very small Δt_0 and Δt_1, is the probability that the gene tree topology is G_1 (that is, $t_0 + t_1 < \tau_1$), that the first coalescent occurs during the time interval $(t_1, t_1 + \Delta t_1)$, and that the second coalescent occurs during the time interval $(t_1 + t_0, t_1 + t_0 + \Delta t_0)$ (see Fig. 1). Equation (9) gives this probability as the sum of four terms. The first term is for the case where the Markov chain is in state 111 right before t_1, with probability $P_{s,111}(t_1)$; the first coalescent occurs in species 1 during $(t_1, t_1 + \Delta t_1)$, with probability

$3\times\frac{2}{\theta_1}\Delta t_1$, the factor 3 due to there being 3 possible pairs for coalescent with the state 111; and then the second coalescent occurs during ($t_1 + t_0$, $t_1 + t_0 + \Delta t_0$) either in population 1, with probability $P_{11,11}(t_0)\times\left(\frac{2}{\theta_1}\Delta t_0\right)$, or in population 2, with probability $P_{11,22}(t_0)\times\left(\frac{2}{\theta_2}\Delta t_0\right)$. Note that in this scenario, the first coalescence changes the state of the chain from 111 to 11. Similarly the 2nd, 3rd, and 4th terms in equation (9) are for the cases where the state right before the first coalescent at time t_1 is 112, 122, and 222, respectively, with the second coalescent occurring either in population 1 or in population 2.

The densities for the other gene trees and for the other initial states are presented in Appendix A and summarized in Tables S3 and S4.

This Markov chain characterization of the genealogical process of coalescent and migration also allows easy calculation of the probabilities of gene tree topologies, integrating over the coalescent times. For example with the initial state '123', the transition probability $P_{123,\,13|23}(\tau_1)$ calculated from the Markov chain of Table 2 (case III) is the probability that sequences 1 and 2 have coalesced by time τ_1. This then gives the probabilities for the five gene trees for the initial state '123' as $P(G_{3c}) = P_{123,\,13|23}(\tau_1)$, $P(G_{6c}) = P(G_{6a}) = P(G_{6b}) = \frac{1}{3}\left(1-P_{123,13|23}(\tau_1)\right)\times \mathrm{e}^{-2/\theta_5(\tau_0-\tau_1)}$, and $P(G_{5c}) = 1 - P(G_{3c}) - 3P(G_{6c})$ (Fig. 1). Here $\mathrm{e}^{-2/\theta_5(\tau_0-\tau_1)}$ is the probability that sequences 1 and 2 do not coalesce in epoch E_2.

In the case of two sequences at a locus, the likelihood calculation given the branch length t is given by equations (7) and (8). The probability density of the genealogy $f(t)$ under M2 (gene flow) is the same as under M_0 for the initial states 13, 23, or 33 (Table 1). For initial states s = 11, 12, or 22, the two sequences can coalesce in any of the three time intervals: (0, τ_1), (τ_1, τ_0), and (τ_0, ∞), so that the density is given as

$$f(t) = \begin{cases} \frac{2}{\theta_1}P_{s,11}(t) + \frac{2}{\theta_2}P_{s,22}(t), & t < \tau_1, \\ \sum\limits_{j\in B_2} P_{s,j}(\tau_1)\times\frac{2}{\theta_5}\mathrm{e}^{-\frac{2}{\theta_5}(t-\tau_1)}, & \tau_1 < t < \tau_0, \\ \sum\limits_{j\in B_2} P_{s,j}(\tau_1)\mathrm{e}^{-\frac{2}{\theta_5}(\tau_0-\tau_1)}\times\frac{2}{\theta_4}\mathrm{e}^{-\frac{2}{\theta_4}(t-\tau_0)}, & t > \tau_0. \end{cases} \qquad (10)$$

where B_2 = {11, 12, 22} is the set of states with two sequences. The transition probability $P_{s,j}(t)$ is calculated using a Markov chain with four states 11, 12, 22, and 1|2. See Hobolth *et al.* (2011).

Likelihood Ratio Test Comparing Models M0 (No Gene Flow) and M2 (gene flow)

As M0 is a special case of M2, we use an LRT to compare them. However, we note that the large-sample χ^2 approximation is not valid and the null distribution (that is, the distribution of the test statistic $2\Delta\ell = 2[\ell_2 - \ell_0]$ when the null hypothesis M0 is true) depends on the data configurations at the loci.

As discussed by Zhu and Yang (2012), if the data consist of loci of configuration 123 only, the symmetric version of model M2 has two more parameters than M0: θ_1 (=θ_2) and M. However, for two reasons, the large-sample χ_2^2 approximation to the test statistic is not valid. First, the null hypothesis M0 corresponds to the alternative hypothesis M2 with $M = 0$, but this parameter value is at the boundary of the parameter space. Second, when $M = 0$, parameter θ_1 (=θ_2) in model M2 becomes unidentifiable. As a result of the violations of the regularity conditions for the χ^2 approximation, the true null distribution is unknown. Furthermore, analysis of data of configuration '123' under M2 leads to an unusual unidentifiability problem: two sets of θ_1 (=θ_2) and M values always give the same log likelihood value.

It is easy to see that this unidentifiability problem exists for the symmetric model if the data consist of a mixture of loci with configurations 12 and 123, or if the 12 and 123 loci are supplemented with an arbitrary mixture of loci of configurations 33, 13, 23, 333, 133, and 233, without any loci of configurations 11, 22, 112, 122, 111, 222, 113, and 223. All such datasets will show the unidentifiability problem under M2 and the two violations of the regularity conditions for the χ_2^2 asymptotics. In this study, we follow Zhu and Yang (2012) and use χ_2^2 as the null distribution to conduct the test and consider the test to be significant if $2\Delta\ell > 5.99$. For data of a mixture of loci with configurations 11, 22, and 12, or of a mixture of 113, 223, and 123, parameter θ_1 (=θ_2) is identifiable in both models M0 and M2. While we still have the problem with the parameter value $M = 0$ at the boundary, the problem is an instance of case 5 in Self and Liang (1987). As a result, the null distribution is known to be the 50:50 mixture of 0 and χ_1^2, with the 5% critical value to be 2.71. The critical values for different mixtures of two initial states under the symmetric model are given in supplementary Table S5.

A similar unidentifiability problem exists under the asymmetrical model for certain combinations of loci. Let $U_1 = \{11, 111, 112, 113\}$ and $U_2 = \{22, 122, 222, 223\}$. If a dataset consists of at least one of the states in U_1 and one of the states in U_2, then M2 is identifiable. In this case, M2 has two more free parameters (M_{12} and M_{21}) than M0 and a 50:50 mixture of 0 and χ_2^2 is the null distribution, with the significance value $2\Delta\ell = 4.61$. If a dataset consists of at least one state in U_1 but none in U_2 or at least one state in U_2 but none in U_1, the model is unidentifiable. In this case the null distribution is unknown and we use χ_3^2 to conduct the test, with critical value 7.82. If a dataset contains none of the states in either U_1 or U_2, we use χ_4^2 to conduct the test, with the critical value 9.49, since M0 and M2 differ by four parameters. The critical values for the likelihood ratio test under the asymmetric model for different mixtures of loci are given in Table S6.

Posterior probabilities of gene tree topologies

When there is gene flow, it may be of interest to know which loci are most likely to have been transferred between species, and to further examine whether the transferred genes share a particular function or are located in the same chromosomal region. Our formulation of the IM model does not allow us to address this question in a straightforward manner. However, we can use an Empirical Bayes approach to calculate the posterior probabilities of the 18 gene tree topologies for each locus, which may be informative about whether the locus is involved in cross-species gene flow. For example, for a '123' locus, the possible gene trees are G_{3c}, G_{5c}, G_{6c}, G_{6a}, and G_{6b}, with G_{3c} being possible only if the locus is transferred between species 1 and 2 (Fig. 1). Similarly for a '112' locus, gene tree shape G_1 is possible only with gene flow. We note that loci of certain configurations, such as '113' or '223', may not provide such information about gene flow.

The probability of data at a locus, $f(D_i | \Theta)$, is a sum over the 18 gene trees (equation 1). The posterior probabilities of the gene trees can be calculated by rescaling those 18 terms so that they sum to 1.

$$f(G_k \mid D_i, \Theta) = \frac{f(G_k \mid \Theta) f(D_i \mid G_k, \Theta)}{f(D_i \mid \Theta)} = \frac{\int_{l_0}^{u_0} \int_{l_1}^{u_1} P(D_i \mid G_k, b_0, b_1) f(G_k, t_0, t_1 \mid \Theta) \, \mathrm{d}t_1 \mathrm{d}t_0}{\sum_k \int_{l_0}^{u_0} \int_{l_1}^{u_1} P(D_i \mid G_k, b_0, b_1) f(G_k, t_0, t_1 \mid \Theta) \, \mathrm{d}t_1 \mathrm{d}t_0}. \qquad (11)$$

We replace the parameters (Θ) by their MLEs $(\hat{\Theta})$, and the method is known as Empirical Bayes (EB). The EB procedure does not account for sampling errors in the MLEs, which may be a concern if the dataset is small and the MLEs involve considerable sampling errors. This is the same EB procedure as used in reconstructing ancestral sequences in molecular phylogenetics (Yang et al., 1995) and in detecting positively selected sites in a protein-coding gene (Nielsen and Yang, 1998).

We conducted a small simulation to examine the reliability of the calculation using equation (11). We simulated datasets using the parameter values: $\tau_0 = 0.0243$, $\tau_1 = 0.0136$, $\theta_4 = 0.0400$, $\theta_5 = 0.0106$, $\theta_1 = 0.0052$, $\theta_2 = 0.0127$, $M_{12} = 0$ and $M_{21} = 0.0183$, which are the MLEs under M2 from the *Drosophila* dataset D1 (auto), to be described and analyzed later (Tables 4 and 9). We simulated two replicate datasets, each of the same size and configurations as the real data. The results are very similar between the two datasets so we discuss only those for the first dataset. The MLEs from the simulated dataset are $\hat{\tau}_0 = 0.0242$, $\hat{\tau}_1 = 0.0137$, $\hat{\theta}_4 = 0.0402$, $\hat{\theta}_5 = 0.0104$, $\hat{\theta}_1 = 0.0058$, $\hat{\theta}_2 = 0.0126$, $\hat{M}_{12} = 0.0018$ and $\hat{M}_{21} = 0.0196$, very close to the true values. The calculated posterior probabilities for gene tree topologies for the '123' loci (Fig. 3a) are accurate in the sense that a posterior probability of 90% is for a correct gene tree about 90% of the time (Fig. 3b). However, the power may not be very high. While the posterior for gene trees G_{6a} and G_{6b} may reach high values, that for G_{6c} is seldom very high (Fig. 3c). It may be hard to distinguish among gene trees G_{3c}, G_{5c}, and G_{6c}. Lastly, approximately equal proportions of loci are inferred to have gene trees G_{6c}, G_{6a} and

G_{6b} (Fig. 3a), and they are also close to the expected proportions. Overall the results indicate a well-behaved method.

Program Implementation, Validation, and Availability

While the general theory of the gene-tree distribution under the Markov chain characterization of the genealogical process under the IM model is straightforward (Zhu and Yang, 2012; Andersen et al., 2014), development of a computer program that can analyze tens of thousands of loci with an arbitrary mixture of loci of different configurations is challenging. Note that under both models M0 (no gene flow) and M2 (gene flow), the number of possible gene trees, the probability density of each gene tree and its coalescent times, and the integration limits for the integrals over the coalescent times all depend on the data configuration at the locus. This dependence makes the programming effort rather tedious and error-prone. Thus we decided to tabulate the necessary results, in Tables S1 and S2 for M0 and similarly in Tables S3 and S4 for M2.

We conducted extensive tests to validate our implementation. The MCCOAL program, which is part of the BPP package (Yang and Rannala, 2010; Zhang et al., 2011), was used to simulate sequence data under models M0 and M2 for different data configurations and parameter values. We ensured consistency of the MLEs: when the same model is used to generate the data and to analyze them, the MLEs should converge to the true parameter values when the size of the dataset (the number of loci) increases. We also confirmed that the likelihood stabilizes when the number of points in the Gaussian quadrature is increased. We simulated 10^6 (true) gene trees under M2 to confirm that the observed frequencies of gene tree topologies match their probabilities calculated from the Markov chain characterization.

Both models M0 and M2 are implemented in the program 3S. We identified two bottlenecks in calculating the likelihood and improved performance in both areas. First, for most initial states, the transition probability matrix $P(t)$ needs to be calculated numerically, involving an expensive matrix exponential. We use the GNU Scientific Library (GSL) (Galassi et al., 2013) to optimize this step. Second, the likelihood calculation is proportional to the number of loci in the data, as it is dominated by the computation of the probability of data at each locus, $f(D_i|\Theta)$. We take advantage of the independence among loci and use OpenMP to parallelize the computation (Dagum and Menon, 1998). While both optimizations are optional, they offer significant speed-ups on genome-scale datasets (Fig. S1). The program, with instructions on how to compile and run it with and without GSL and OpenMP, is available at http://abacus.gene.ucl.ac.uk/software/3s.html.

Drosophila *genomic datasets*

We compiled multi-locus datasets for three *Drosophila* species, *D. melanogaster* (M), *D. simulans* (S) and *D. yakuba* (Y). We used Flybase FB2016_01 (Attrill et al., 2016) genome releases of *D. melanogaster* (r6.09, January 2016), *D. simulans* (r2.01, Hu et al., 2013) and *D. yakuba* (r1.05, January 2016), as well as the assembly of *D. simulans* strain M252 (Palmieri et al., 2014). We treated the two *D. simulans* genomes (r2.01 from North American and M252 from Madagascar) as two random samples from the same species. Five datasets of MSSY loci were constructed (Table 4): D1 (auto) for autosomes 2 and 3, D2 (noncoding) for intergenic regions and introns from chromosomes 2 and 3, D3 (chrX) for the X chromosome, D4 (exons complete) and D5 (exons split). D4 (exons complete) was compiled using non-overlapping complete exons on chromosomes 2 and 3. When exons were overlapping, only the longest was kept. For all datasets except D4 (exons complete), sequences were split into chunks between 100 and 500bp that were separated by at least 2kb. These criteria were from Wang and Hey (2010), based on previous estimates of recombination rates for *Drosophila* (Hey and Nielsen, 2004). To construct each of datasets D1-D4, we extracted the loci from the *D. melanogaster* genome as a starting point and then ran NCBI BLAST (Camacho et al., 2009) with default settings to find matching sequences in the other genomes. We discarded short matches (<40% of the query sequence), and removed loci where the two longest matches differed in length by less than 10% to avoid paralogues. The remaining loci were aligned using MAFFT, using default settings (Katoh and Standley, 2013). We reduced each of the MSSY loci to either MSY or SSY by randomly removing either the *D. melanogaster* or one of the *D. simulans* sequences. Dataset D5 (exons split) was constructed by splitting the alignments of D4 (exons complete) into loci of between 100 and 500bp and removing chunks that did not fulfill the 2kb-separation criterion. Thus all loci in D5 are also in D4, but the alignments of the same loci in D5 may be shorter. Some loci in D4 (374 of them) were longer than 2600bp, and were split into more than one locus in D5. Finally, we added the 378 MMY loci from Hutter et al. (2007) to all datasets except D2 (chrX) after updating their coordinates to the current *D. melanogaster* release and confirming that they do not overlap with the MSSY loci we compiled.

Note that D2 (noncoding) includes both intergenic regions and introns: these were found to produce very similar estimates in a preliminary analysis and were thus merged into one dataset. D1 (auto) and D3 (chrX) include both noncoding regions and exons. The loci in D2 (noncoding), D4 (exons complete), and D5 (exons split) may not be included in D1 (auto).

The five datasets were analyzed using the program 3S under models M0 and M2 to estimate parameters and to test for gene flow. Fitting the two models to each dataset took about 20 minutes on a single core and ~1 minute using 32 cores on a Sun Fire X4600M2 server (with 32 Opteron AMD cores at 2.7GHz). We also calculated the posterior probabilities of gene tree topologies under M2 to identify the gene loci that are most likely to have been transferred across species barriers during introgression (Eq. 11).

RESULTS

Computer Simulation to Examine the Statistical Properties of the new model

We conducted computer simulations to examine the false positive rate and the power of the LRT comparing models M0 (no gene flow) and M2 (gene flow) to test for migration between species 1 and

2. We also examined the biases and variances of MLEs of parameters under M2. Our simulation design largely follows that of Zhu and Yang (2012).

To examine the false positive rate of the test, we simulated replicate datasets under the symmetrical version of M0 and analyzed them under both M0 and M2, assuming symmetry (Table 5). We used four sets of parameter values (Zhu and Yang, 2012: table 1). The first two sets are based roughly on parameter 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 number of loci was fixed at L = 10, 100, 1000, and 15,000, with each locus having 500 sites. Gene trees with branch lengths (coalescent times) were generated from the multispecies coalescent model (Rannala and Yang, 2003) using the program MCCOAL, which is part of the BPP pacakge (Rannala and Yang, 2003; Yang and Rannala, 2010). Given the gene tree, the sequences were allowed to evolve along the branches of the tree, under the JC69 mutation model (Jukes and Cantor, 1969). The resulting sequences at the tips of the tree constituted the data. Each replicate dataset thus consisted of L sequence alignments, with 500 base pairs at each locus. We considered three kinds of data: (a) all loci of configuration 123, (b) a mixture of loci of configurations 11 and 12 in equal proportions, and (c) a mixture of loci of configurations 113 and 123 in equal proportions. The number of replicates was 1000.

Overall, the use of the χ_2^2 distribution for data of configuration (a) 123 made the test conservative, as the false positive rate was always <1%, while an error rate of 5% was allowed (Table 5). For the 'pairs' data (configuration b, 11&12), we observed false positive rates of up to10% for parameter sets 2 and 3. The analysis seemed to suffer from a lack of information when only two sequences were available at each locus. In theory the false positive rate should converge to 5% when the number of loci increases, so it appears that more loci are needed for the asymptotics to be reliable for the 'pairs' data than for the 'triplet' data (c: 113&123). Adding an outgroup sequence increased the information content in the data, reducing the false positive rate to below 5%.

We examined the power of the test by simulating sequence alignments under the symmetrical version of M2 (gene flow). We used parameter values of Set 1 (hominoid) and Set 2 (mangroves), with $M_{12} = M_{21} = 1$ (Table 6). The test has virtually no power with L = 10 loci. With L = 100 or 1000, there are large performance differences between the two sets of parameter values. This is because the sequences are far more divergent and thus more informative for the mangroves set than for the hominoid set. Power is quite high with 1000 loci, when three sequences are used at each locus. Power is similar for the '123' data and for the '113&123' data. There is dramatic difference in power between the 'pairs' data (b, 11&12) and the 'triplet' data (c, 113&123). The use of the outgroup species improves the power of the test dramatically. This is consistent with Lohse *et al.* (2011), who suggested that triplet samples provide qualitatively new information about historical parameters in the joint distribution of topologies and branch lengths.

Table 7 lists the means and standard deviations of the MLEs of parameters under model M2 for the same data analyzed in Table 6. Datasets with '123' loci only suffer from the problem of unidentifiability and do not allow the estimation of the migration rate. Inclusion of the '113' loci allows the model to estimate θ_1 ($=\theta_2$) and M and the unidentifiability problem disappears, leading to better parameter estimation. Furthermore the 'triplet' data provided much better parameter estimates than the 'pair' data.

We also simulated data under the general (asymmetrical) model M2 (gene flow) to examine the estimation of migration rates. Given that the estimation was poor for the 'pair' data even under the symmetrical model (Table 7) and that the asymmetrical model involves even more parameters, we focus on the 'triplet' data only, with three sequences per locus. We used the mangrove set of parameters, with the migration rates set at $M_{12} = 0.1$ and $M_{21} = 1$ migrant individuals per generation. We explored two different data configurations, with each dataset consisting of (a) '223' and '123' loci in equal proportions, and (b) '113', '223', and '123' loci in equal proportions (Table 8). The results suggest that 100 loci may be too few to obtain reliable parameter estimates. In particular, the lack of polymorphism data for species 1 in the 223&123 configuration led to large fluctuations in the estimates of θ_5, θ_1 and M_{21}. Even with 1000 loci, we encountered several datasets in which the MLEs of parameters hit the boundary set in the program (with $M_{12} = M_{21} = 0$), or the MLEs imply a star tree (with $\tau_0 \approx \tau_1$ and $\theta_5 \approx 0$ or ∞). With 15000 loci, the estimates are close to the true values. Estimates of migration rates are seen to involve a positive bias, but the bias is small with 15000 loci. To fit the asymmetrical IM model, it appears important to include thousands of loci, and to include population data for both species 1 and 2 (such as '113' and '223' loci), as well as the '123' loci.

Analysis of Drosophila *genomic datasets*

For each of the five datasets (Table 4), we performed three runs of 3S and used the results from the run with the highest log likelihood. Integration over coalescent times in the gene trees used Gaussian quadrature with K =16 points. We used both the symmetrical and asymmetrical versions of models M0 and M2, but here we focus on the asymmetrical models as they fit the data much better (Table 9). We describe some general features of the results before discussing results specific to individual datasets. In every dataset, the LRT comparing M0 and M2 is significant. Furthermore, the parameter estimates under M2 suggest no migration from *D. melanogaster* to *D. simulans*, and about 0.016 to 0.044 immigrants per generation from *D. simulans* to *D. melanogaster*. The consistency among the datasets suggests that this pattern of unidirectional migration may be real. Estimates of τ and θ parameters have very small standard errors because of the large size of the datasets. Parameter estimates are nearly identical between datasets D1 (auto) and D2 (noncoding), and between D4 (exons complete) and D5 (exons split), suggesting that with such large genomic datasets, how extensively the genomes were sampled to compile the datasets did not matter much. Note that the autosomal dataset

D1 (auto) is dominated by noncoding DNA, even though different noncoding loci may be included in D1 and D2, and that loci in D5 (exons split) are a subset of those in D4 (exons complete). While model M0 did not fit the data as well as M2, it produced stable and reasonable estimates of θ and τ parameters, which were also similar to estimates from M2. (The exon datasets D4 and D5 are exceptions to this pattern, to be discussed later.) For example, in datasets D1 (auto) and D2 (noncoding), both M0 and M2 estimates suggest that θ_S (≈0.013) is more than twice as large as θ_M (≈0.005-0.006), consistent with previous studies which suggest that *D. simulans* has a larger effective population size than *D. melanogaster* (e.g., Langley et al., 2012; Wang and Hey, 2010). Also from datasets D1 (auto) and D2 (noncoding) we obtained $\hat{\tau}_{MS}$ = 0.011 and $\hat{\theta}_{MS}$ = 0.013-0.014 under M0, and $\hat{\tau}_{MS}$ = 0.012-0.014 and $\hat{\theta}_{MS}$ = 0.011-0.012 under M2 (Table 9). The slightly smaller estimates of τ_{MS} and larger estimates of θ_M under M0 than under M2 may be expected because a more recent divergence between *D. melanogaster* and *D. simulans* and a larger population size for *D. melanogaster* may help M0 (which does not allow gene flow) to explain the genetic variation introduced by immigrants from *D. simulans*.

Dataset D3 (chrX) for the X chromosome showed very different patterns from the autosomal datasets D1 (auto) and D2 (noncoding), with a smaller estimate of θ_S, and slightly larger estimates of the other θ parameters. The estimated migration rate M_{SM} was much higher for the X than for the autosomes. By the simple model of random mating and neutral evolution, and assuming the same mutation rate for the X and the autosomes, one would expect the effective population size for the X chromosome to be ¾ that for the autosome, so that θs for X should be ¾ times as large as θs for the autosomes, while the τs and Ms should be identical. The parameter estimates suggested that this simplistic model may not fit the data well. However the estimates of θ_M and M_{SM} from D3 (chrX) were associated with large sampling errors. Indeed D3 (chrX) does not include any MMY loci, so that the data contain only very weak information concerning θ_M even though the model is identifiable. The correlation between estimates of θ_M and M_{SM} means that estimation of M_{SM} may be affected as well. We thus reran M2 under the constraint that $\theta_M = ½\theta_S$ or $\theta_M = \theta_S$, obtaining estimates of M_{SM} to be 0.016 and 0.008 (Table 9). Thus there was no evidence for a large M_{SM} for the X than for the autosomes. The large changes to θ_M and M_{SM} caused virtually no change to the log likelihood or to estimates of other parameters, suggesting that the data are uninformative about θ_M and M_{SM} while the other parameters were well estimated. We leave it to future investigations, perhaps by including some MMY or MMM loci with polymorphism for *D. melanogaster*, to generate more reliable parameter estimates for the X and to understand possible differences in the evolutionary process between the X chromosome and the autosomes.

The two exon datasets, D4 (exons complete) and D5 (exons split), are exceptional to the general pattern of high similarity of parameter estimates between M0 and M2. For those two datasets, estimates of τ_{MS} under M2 are much larger than those under M0. However those M2 estimates are unreliable, because ML optimization under M2 converged to a star tree with $\tau_{\mathrm{MSY}} \approx \tau_{\mathrm{MS}}$ and $\theta_{\mathrm{MS}} \approx 0$ (Table 9). We were unable to determine the reasons for this behavior. We note that the same behavior was encountered in a few simulated datasets, as mentioned earlier, and that the problem did not occur for dataset D1 (auto), which includes both coding and non-coding loci. The estimates of θ_{M} and θ_{S} from D4 (exons complete) and D5 (exons split) were smaller than those from D1 (auto) or D2 (noncoding), which can be explained by the reduced neutral mutation rate in the exons due to selective constraint on nonsynonymous mutations. Again, the estimates suggest no migration from *D. melanogaster* to *D. simulans*, but the migration rate from *D. simulans* to *D. melanogaster* is much higher than for the autosome. We note that estimates of τ and θ parameters under M0 from those exon datasets were similar to the M0 estimates from D1 (auto) and D2 (non-coding), and that the estimates of τ_{MSY} were very similar between M0 and M2 for the same dataset. Thus we reran the M2 analysis of the two exon datasets, with $\tau_{\mathrm{MSY}} = 0.020$ and $\tau_{\mathrm{MS}} = 0.013$ fixed, to estimate the other parameters. The results appear much more reasonable (Table 9). Both datasets D4 and D5 suggested no migration from *D. melanogaster* to *D. simulans*, but the estimates of M_{SM}, at ~0.02 immigrants from *D. simulans* to *D. melanogaster* per generation, were very similar to those from D1 (auto) and D2 (noncoding).

To examine the robustness of our estimates of migration rates and to explore the impact of the correlation between population sizes and migration rates, we re-analyzed the datasets under M2 (gene flow) assuming asymmetrical migration rates (with $M_{\mathrm{MS}} \neq M_{\mathrm{SM}}$) but symmetrical population sizes ($\theta_{\mathrm{M}} = \theta_{\mathrm{S}}$) (Table S7). Again the LRT is significant in every dataset, and parameter estimates suggested unidirectional migration, with $\hat{M}_{\mathrm{MS}} = 0$ in every dataset. However, estimates of M_{SM} were much larger than those of Table 9 in every dataset except for D3 (chrX), which has been discussed above. For example, $\hat{M}_{\mathrm{SM}} = 0.036\text{-}0.041$ from D1 (auto) and D2 (noncoding) under the constraint $\theta_{\mathrm{M}} = \theta_{\mathrm{S}}$ (Table S7), in comparison with 0.016-0.018 without the constraint (Table 9). We note that, except for θ_{M} and M_{SM}, the parameter estimates were virtually identical with and without the constraint $\theta_{\mathrm{M}} = \theta_{\mathrm{S}}$ (compare Tables S7 and 9). There are far more SSY than MMY loci in those datasets (Table 4), so that the estimates of $\theta_{\mathrm{M}} = \theta_{\mathrm{S}}$, at 0.012 (Table S7), were dominated by the *D. simulans* polymorphism data, and were too large for *D. melanogaster*. This has lead to overestimates of M_{SM}, apparently because a large M_{SM} is more compatible with the (unrealistically assumed) large θ_{M}. Thus the assumption $\theta_{\mathrm{M}} = \theta_{\mathrm{S}}$ has caused serious biases in the estimation of migration rates, highlighting the importance of the asymmetrical model. Note that the data contain strong evidence against the assumption $\theta_{\mathrm{M}} = \theta_{\mathrm{S}}$; for example, relaxing the assumption improves the log likelihood by 66-82 units in datasets D1 (auto) and D2 (noncoding). D3 (chrX) does not include any MMY loci. As a result, θ_{M} is unidentifiable under M0 (so that the log likelihood is the same with and without the constraint

$\theta_M = \theta_S$), while under M2, θ_M is identifiable but very poorly estimated (so that the log likelihoods are distinct but extremely similar with and without the constraint) (Tables 9 and S7).

We used equation (11) to calculate the posterior probabilities for gene trees for the MSY loci in the five datasets (Table 4). Here we discuss the results for D5 (exons split) (Fig. 4), and those for D1 (auto) and D3 (chrX) are presented in Figs. S2 and S3. At the MLEs under M2 (Table 9, with τ_{MSY} = 0.020 and τ_{MS} = 0.013 fixed), the expected gene tree probabilities for any MSY locus are $P(G_{3c})$ = 0.1324, $P(G_{5c})$ = 0.7368, and $P(G_{6c}) = P(G_{6a}) = P(G_{6b})$ = 0.0436, with the gene tree-species tree mismatch probability $P(G_{6a}) + P(G_{6b})$ = 0.0872. Most loci have gene tree G_{5c} (Fig. 4), because the migration rate is low, so that G_{3c} is uncommon and because the outgroup species is quite distant so that there is not much gene tree-species tree discordance. A small proportion of loci very likely have the gene tree G_{3c}, and are likely to have been transferred across species (from *D. simulans* to *D. melanogaster* since $M_{MS} \approx 0$). The top 41 loci, with $P(G_{3c})$ > 95%, are listed in Table S8. More than half of those loci were also inferred to have $P(G_{3c})$ > 95% in the analysis of dataset D4 (exons complete) (Table S8), suggesting that this inference was not very sensitive to the different filtering procedures applied to compile the datasets.

An intriguing feature in Fig. 4 (and also in Figs. S2 and S3 for datasets D1 and D3) is that many more loci seem to support gene tree G_{6c} than G_{6a} or G_{6b}, while the model predicts equal proportions for those three gene trees. This is in contrast to the simulated dataset, in which the three gene trees are inferred to occur with similar proportions, as expected under the model (Fig. 3A). The reasons for this pattern are unknown, but are likely to be some kind of model violation.

To explore the potential of the IM model for species tree estimation under the multispecies coalescent with migration, we applied model M2 to dataset D1 (auto), assuming alternative species trees for M, S, and Y. The MLEs and log likelihood values are shown in Table 10. The ((MS)Y) tree has a much greater log likelihood value than the two alternative trees (by about 20,000 units). Indeed, both alternative trees converge to the star tree with $\tau_0 = \tau_1$. Migration is detected only in the direction of S→M when the assumed tree is ((MS)Y). Note that our model assumes migration between the two ingroup species only. In theory a stratified bootstrap resampling procedure can be used to assess the significance of the ML species tree, sampling loci and then sampling sites for each sampled locus. This is not pursued here since there does not seem to be any uncertainty about the species phylogeny in this case (Russo et al., 1995; Obbard et al., 2012).

DISCUSSION

Utilities and limitations of our implementation

In this paper, we have extended our previous implementation of the IM model (Zhu and Yang, 2012) in several important ways. First, we have relaxed the symmetry assumption, so that the test of gene flow and estimation of migration rates and population size parameters can be conducted under more realistic models. For the *Drosophila* datasets, our analyses suggest that gene flow is indeed asymmetrical, the population sizes of *D. melanogaster* and *D. simulans* are very different, and accounting for such asymmetries in the model is important to accurate estimation of the migration rates. Second, we have extended the implementation so that a locus can have 2 or 3 sequences of arbitrary configurations. This removes the unidentifiability problem that we encountered when '123' loci alone were used, making it possible to estimate the migration rates. It also improves the power of the LRT of gene flow because the null distribution becomes known. The extension to arbitrary loci also paves the way for implementing more complex models of migration.

We envisage that a major future use of the IM model is to infer species phylogenies under the multispecies coalescent model with migration, accommodating two major factors that thwart species tree estimation, especially for species formed during radiative speciations: incomplete lineage sorting (ILS) and gene flow (Mallet et al., 2016). Heuristic methods based on the model that treat estimated gene tree topologies as observed data are being developed (Wen et al., 2016), but full likelihood methods have the advantage of accommodating the different sources of uncertainties appropriately. However the functionality of 3S in this regard is limited. The assumption of gene flow between sister species only may be too restrictive and gene flow between non-sister species needs to be allowed as well (Mallet et al., 2016). Furthermore, our implementation is restricted to three species, with two or three sequences per locus. This limitation is mainly due to our use of numerical integration (Gaussian quadrature) to integrate over the coalescent times, with the dimension of the integrals to be one less than the number of sequences at the locus. With four or more sequences per locus, this calculation may not be feasible. Furthermore, the number of states in the Markov chain used to characterize the genealogical process also increases explosively with the increase of the number of sequences per locus (Andersen et al., 2014). We suggest that to analyze genomic datasets involving more than three species and more than three sequences per locus, a subsampling procedure may be useful, similarly to our analysis of the *Drosophila* datasets (see also Wang and Hey, 2010). Suppose there are $s > 3$ species. We specify a 'master' species tree including all s species and use it to define the parameters: the $(s - 1)$ species divergence times (τs) and up to $(2s - 1)$ population size parameters (θs). At every locus, we sample three sequences, which may be from different species, so that the data configurations may be 123, 114, 255, etc. The species tree for the sequences of any particular locus can be constructed from the master species tree by pruning off branches for species not sampled in the data at the locus. The theory developed in Zhu and Yang (2012) and in this paper will then be applicable with the only complication that the coalescent rate (the population size) and the migration rate may change along the same branch on the species subtree at the speciation events in the master species tree. Such rate changes are relatively straightforward to accommodate. This strategy involves filtering of data but the information loss may not be very serious for such large genomic datasets. Note that given the data, this strategy calculates the likelihood correctly.

In the future, we also hope to implement models of nonhomogeneous migration rates over time. Gene flow may be common at the early stage of species formation and decrease until the two populations achieve complete isolation. A simple model may assume a constant migration rate M since species divergence until a time point T ($0 < T < \tau_1$) when gene flow ceases. In this model of *isolation with initial migration*, both the migration rate M and the time point T will be parameters to be estimated from the sequence data (Wilkinson-Herbots, 2012). The same Markov chain characterization as used here can be used to derive the density of gene trees by breaking the time epoch E_1 into two segments: E_{1a}: $0 < t < T$ and E_{1b}: $T < t < \tau_1$. Alternatively, one may use a deterministic mathematical function such as an exponential decay to describe the changing migration rate over time. The initial migration rate and the exponential decay rate will be parameters to be estimated. If reproductive isolation builds up gradually after species split, such nonhomogeneous migration models may be more realistic than the usual IM model with a constant migration rate after species divergence.

Similarly, introgression or hybridisation may be modelled in the same framework (Twyford and Ennos, 2011). Recent introgression or contamination may be modelled by assuming that a proportion of individuals sampled from species 1 are in fact from species 2. Introgression can then be tested using a likelihood ratio test. As the model naturally accommodates ancestral polymorphism and incomplete lineage sorting (ILS), the test will distinguish introgression from ILS. Note that introgression affects all loci of the introgressed individual, while with ILS, caused by the coalescent process, the different genomic loci have independent histories.

Asymmetrical Migration in Drosophila *fruit flies*

Wang and Hey (2010: Table 7) compiled and analyzed a *Drosophila* dataset similar to our dataset D1 (auto), consisting of 30,323 autosomal loci but including only two sequences for each locus, of configurations SS, MS, and MM. Under the asymmetrical model, their estimates of population size parameters are $\theta_\text{M} = 0.0055$ and $\theta_\text{S} = 0.01352$, which are close to our estimates from D1 (auto). The ancestral population size θ_MS estimated by Wang and Hey ranges from 0.007 to 0.010, whereas our estimates are larger, at $\theta_\text{MS} = 0.011$ and $\theta_\text{MSY} = 0.040$. The M-S divergence time parameter is $\tau_\text{MS} = 0.017$ by Wang and Hey and 0.0136 in our analysis. A strong negative correlation between τ_MS and θ_MS is expected in such analyses (Yang, 2002). Wang and Hey (2010) estimated the migration rate (in our notation) to be $M_\text{MS} = N_\text{S}m_\text{MS} = 0$ from *D. melanogaster* to D. *simulans* and $M_\text{SM} = N_\text{M}m_\text{SM} = 4.846 \times 0.00552/4 = 0.0067$ from *simulans* to *melanogaster*. Our estimates under M2 are $M_\text{MS} = 0$ as well and $M_\text{SM} = 0.0183$, which is much larger.

The data of Wang and Hey (2010) were also analyzed by Lohse *et al.* (2011, Table 1), who compared parameter estimates from two datasets which have either two or three sequences per locus for the same set of loci. The authors found that the estimate of the migration rate from the ‘triplet’ data was nearly twice as large as that for the ‘pair’ data. This is consistent with our finding.

We note that our datasets are based on updated genome sequences, relative to the data analyzed by Wang and Hey (2010) and Lohse *et al.* (2011). Also different filters were applied and different loci were included in those datasets. Furthermore, Wang and Hey (2010) removed loci at which the pairwise sequence distances indicated gene tree-species tree conflict. We did not apply this filtering because such loci are informative about the gene tree distribution and about the parameters in our analysis of loci of three sequences. Lohse *et al.* (2011) removed highly variable loci and highly variable sites so that the data could be analyzed under the infinite-sites model. Given the multiple differences among the datasets, we conclude that the estimates obtained from those studies are largely consistent.

Different from Wang and Hey (2010), we also compiled and analyzed a dataset for the X chromosome (D3 chrX) as well as two exon datasets: D4 (exons complete) and D5 (exons split). The use of multiple datasets, even though some of them are overlapping, allows us to confirm the robustness of our analyses, as processes such as migration are expected to have genome-wide effects, and to discover similarities and differences in the evolutionary process among different parts of the genome. Indeed all five datasets we analyzed support a model of unidirectional gene flow, from *D. simulans* to *D. melanogaster*, at the rate of ~0.02 migrant individuals per generation. We included the two exon datasets even though we do not expect exons to be evolving neutrally. Note that the multispecies coalescent model implemented in 3S assumes neutral evolution of the gene sequences, such that mutations in the sequences do not affect the genealogical process or the gene tree distribution. Nevertheless, most proteins appear to perform the same conserved function in closely related species and their coding genes are under similar purifying selection in the different species. The main effect of the selective constraint may then be a reduction of the neutral mutation rate. Species-specific natural selection such as positive selection would be more problematic but loci undergoing positive selection or responsible for between-species incompatibilities are expected to be rare. Similar points have been made by Ebersberger et al. (2007; see also Yang, 2015) in their analysis of hominoid genomic sequence data.

Acknowledgments. We thank Thomas Buckley, Bastien Boussau, and an anonymous reviewer for many critical and constructive comments, which have led to improvement of our ms. We thank Bastien Boussau for the suggestion of inferring the gene loci that may have been transferred across species due to gene flow. This study is supported by a grant from the Biotechnological and Biological Sciences Research Council (BBSRC) to Z.Y. T.Z. is supported by Natural Science Foundation grants (31301093, 11301294 and 11201224), and a grant from the Youth Innovation Promotion Association of Chinese Academy of Sciences (2015080).

REFERENCES

Andersen, L. N., T. Mailund, and A. Hobolth. 2014. Efficient computation in the IM model. J. Math. Biol. 68:1423-1451.

Attrill, H., K. Falls, J. L. Goodman, G. H. Millburn, G. Antonazzo, A. J. Rey, and S. J. Marygold. 2016. FlyBase: establishing a Gene Group resource for *Drosophila melanogaster*. Nucleic Acids Res. 44:D786-792.

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

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

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

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

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

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Chan, Y. C., C. Roos, M. Inoue-Murayama, E. Inoue, C. C. Shih, K. J. Pei, and L. Vigilant. 2013. Inferring the evolutionary histories of divergences in Hylobates and Nomascus gibbons through multilocus sequence data. BMC Evol. Biol. 13:82.

Dagum, L., and R. Menon. 1998. OpenMP: an industry standard API for shared-memory programming. Computational Science & Engineering, IEEE5 1:46-55.

Ebersberger, I., P. Galgoczy, S. Taudien, S. Taenzer, M. Platzer, and A. von Haeseler. 2007. Mapping human genetic ancestry. Mol. Biol. Evol. 24:2266-2276.

Edwards, S. V. 2009. Is a new and general theory of molecular systematics emerging? Evolution 63:1–19.

Ellegren, H., L. Smeds, R. Burri, P. I. Olason, N. Backstrom, T. Kawakami, A. Kunstner, H. Makinen, K. Nadachowska-Brzyska, A. Qvarnstrom, S. Uebbing, and J. B. W. Wolf. 2012. The genomic landscape of species divergence in Ficedula flycatchers. Nature 491:756-760.

Fontaine, M. C., J. B. Pease, A. Steele, R. M. Waterhouse, D. E. Neafsey, I. V. Sharakhov, X. Jiang, A. B. Hall, F. Catteruccia, E. Kakani, S. N. Mitchell, Y. C. Wu, H. A. Smith, R. R. Love, M. K. Lawniczak, M. A. Slotman, S. J. Emrich, M. W. Hahn, and N. J. Besansky. 2015. Mosquito genomics. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science 347:1258524.

Galassi, M., J. Davies, J. Theiler, B. Gough, R. Priedhorsky, G. Jungman, and M. Booth. 2013. *GNU Scientific Library Reference Manual*. The GSL Project.

Gronau, I., M. J. Hubisz, B. Gulko, C. G. Danko, and A. Siepel. 2011. Bayesian inference of ancient human demography from individual genome sequences. Nature Genet. 43:1031-1034.

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

Hey, J., and R. Nielsen. 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., L. N. Andersen, and T. Mailund. 2011. On computing the coalescence time density in an isolation-with-migration model with few samples. Genetics 187:1241-1243.

Hu, T. T., M. B. Eisen, K. R. Thornton, and P. Andolfatto. 2013. A second-generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. Genome Res. 23:89-98.

Hutter, S., H. Li, S. Beisswanger, D. De Lorenzo, and W. Stephan. 2007. Distinctly different sex ratios in African and European populations of *Drosophila melanogaster* inferred from chromosomewide single nucleotide polymorphism data. Genetics 177:469-480.

Innan, H., and H. Watanabe. 2006. The effect of gene flow on the coalescent time in the human-chimpanzee ancestral population. Mol. Biol. Evol. 23:1040-1047.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21-123 *in* Mammalian Protein Metabolism (H. N. Munro, ed.) Academic Press, New York.

Katoh, K., and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772-780.

Kimura, M., and W. H. Weiss. 1964. The stepping stone model of genetic structure and the decrease of genetic correlation with distance. Genetics 49:561-576.

Kingman, J. F. C. 1982. The coalescent. Stochastic Process Appl. 13:235-248.

Kutschera, V. E., T. Bidon, F. Hailer, J. L. Rodi, S. R. Fain, and A. Janke. 2014. Bears in a forest of gene trees: Phylogenetic inference is complicated by incomplete lineage sorting and gene flow. Mol. Biol. Evol. 31:2004-2017.

Langley, C. H., K. Stevens, C. Cardeno, Y. C. Lee, D. R. Schrider, J. E. Pool, S. A. Langley, C. Suarez, R. B. Corbett-Detig, B. Kolaczkowski, S. Fang, P. M. Nista, A. K. Holloway, A. D. Kern, C. N. Dewey, Y. S. Song, M. W. Hahn, and D. J. Begun. 2012. Genomic variation in natural populations of *Drosophila melanogaster*. Genetics 192:533-598.

Leaché, A. D., R. B. Harris, M. E. Maliska, and C. W. Linkem. 2013. Comparative species divergence across eight triplets of spiny lizards (Sceloporus) using genomic sequence data. Genome Biol. Evol. 5:2410-2419.

Li, W.-H. 1976. Distribution of nucleotide differences betwen two randomly chosen cistrons in a subdivided population: the finite island model. Theor Popul Biol 10:303-308.

Liu, S., E. D. Lorenzen, M. Fumagalli, B. Li, K. Harris, Z. Xiong, L. Zhou, T. S. Korneliussen, M. Somel, C. Babbitt, G. Wray, J. Li, W. He, Z. Wang, W. Fu, X. Xiang, C. C. Morgan, A. Doherty, M. J. O'Connell, J. O. McInerney, E. W. Born, L. Dalen, R. Dietz, L. Orlando, C. Sonne, G. Zhang, R. Nielsen, E. Willerslev, and J. Wang. 2014. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. Cell 157:785-794.

Lohse, K., R. J. Harrison, and N. H. Barton. 2011. A general method for calculating likelihoods under the coalescent process. Genetics 189:977-987.

Mallet, J. 2005. Hybridization as an invasion of the genome. Trends Ecol. Evol. 20:229-237.

Mallet, J., N. Besansky, and M. W. Hahn. 2016. How reticulated are species? BioEssays.

Martin, S. H., K. K. Dasmahapatra, N. J. Nadeau, C. Salazar, J. R. Walters, F. Simpson, M. Blaxter, A. Manica, J. Mallet, and C. D. Jiggins. 2013. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Res 23:1817-1828.

Melo-Ferreira, J., P. Boursot, M. Carneiro, P. J. Esteves, L. Farelo, and P. C. Alves. 2012. Recurrent introgression of mitochondrial DNA among hares (Lepus spp.) revealed by species-tree inference and coalescent simulations. Syst. Biol. 61:367-381.

Nath, H. B., and R. C. Griffiths. 1993. The coalescent in two colonies with symmetric migration. J. Math. Biol. 31:841–852.

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

Nielsen, R., and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929-936.

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

Obbard, D. J., J. Maclennan, K. W. Kim, A. Rambaut, P. M. O'Grady, and F. M. Jiggins. 2012. Estimating divergence dates and substitution rates in the *Drosophila* phylogeny. Mol. Biol. Evol. 29:3459-3473.

Palmieri, N., V. Nolte, J. Chen, and C. Schlotterer. 2014. Genome assembly and annotation of a *Drosophila simulans* strain from Madagascar. Mol. Ecol. Resour. 15:372-381.

Patterson, N., D. J. Richter, S. Gnerre, E. S. Lander, and D. Reich. 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature 441:1103-1108.

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

Russo, C. A., N. Takezaki, and M. Nei. 1995. Molecular phylogeny and divergence times of

Drosophilid species. Mol. Biol. Evol. 12:391-404.

Saitou, N. 1988. Property and efficiency of the maximum likelihood method for molecular phylogeny. J. Mol. Evol. 27:261-273.

Self, S. G., and K.-Y. Liang. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J. Am. Stat. Assoc. 82:605-610.

Strobeck, K. 1987. Average number of nucleotide differences in a sample from a single subpopulation: A test for population subdivision. Genetics 117:149-153.

Takahata, N. 1988. The coalescent in two partially isolated diffusion populations. Genet. Res. (Camb.) 52:213-222.

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

Twyford, A. D., and R. A. Ennos. 2011. Next-generation hybridization and introgression. Heredity 108:179-189.

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

Wen, D., Y. Yu, M. W. Hahn, and L. Nakhleh. 2016. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. Mol. Ecol.

Wilkinson-Herbots, H. M. 1998. Genealogy and subpopulation differentiation under various models of population structure. J. Math. Biol. 37:535-585.

Wilkinson-Herbots, H. M. 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.

Wilkinson-Herbots, H. M. 2012. The distribution of the coalescence time and the number of pairwise nucleotide differences in a model of population divergence or speciation with an initial period of gene flow. Theor. Popul. Biol. 82:92-108.

Wright, S. 1931. Evolution in Mendelian populations. Genetics 16:97-159.

Wright, S. 1943. Isolation by distance. Genetics 28:114-138.

Yamamichi, M., J. Gojobori, and H. Innan. 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. 2002. Likelihood and Bayes estimation of ancestral population sizes in Hominoids using data from multiple loci. Genetics 162:1811-1823.

Yang, Z. 2010. A likelihood ratio test of speciation with gene flow using genomic sequence data. Genom. Biol. Evol. 2:200-211.

Yang, Z. 2014. *Molecular Evolution: A Statistical Approach*. Oxford University Press, Oxford, England.

Yang, Z. 2015. The BPP program for species tree estimation and species delimitation. Curr. Zool. 61:854-865.

Yang, Z., S. Kumar, and M. Nei. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641-1650.

Yang, Z., and B. Rannala. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl. Acad. Sci. U.S.A. 107:9264-9269.

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

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

Zhou, W. W., Y. Wen, J. Fu, Y. B. Xu, J. Q. Jin, L. Ding, M. S. Min, J. Che, and Y. P. Zhang. 2012. Speciation in the *Rana chensinensis* species complex and its relationship to the uplift of the Qinghai-Tibetan Plateau. Mol. Ecol. 21:960-973.

Zhu, T., and Z. Yang. 2012. Maximum likelihood implementation of an isolation-with-migration model with three species for testing speciation with gene flow. Mol. Biol. Evol. 29:3131-3142.

Appendix A.

Distribution of gene trees for three sequences under M2 (gene flow)

Case I: Initial states 111 and 222

With the initial state s = 111 or 222, all three sequences at the locus are from the same species (1 or 2). Due to the symmetry, the densities of the three gene trees of the same shape (such as G_{1c}, G_{1a}, and G_{1b}) are identical. There is thus no need to keep track of the sequence IDs, even though the likelihood averages over all 18 gene trees (Table S1). Thus we consider a Markov chain with 8 states: 111, 112, 122, 222, 11, 12, 22, 1|2, with '1|2' to be an artificial state formed by merging states 1 and 2. The rate matrix is given in Table 3. The density for gene tree shape G_1 is given in equation (9). By a similar argument we obtain the densities for tree shapes G_2-G_6, as follows.

$$\begin{aligned}
f(G_2, t_0, t_1) &= \tfrac{2}{\theta_5} \mathrm{e}^{-\frac{2}{\theta_5} t_0} \times \\
&\sum_{j \in S_2} \Big[3\tfrac{2}{\theta_1} P_{s,111}(t_1) P_{11,j}(\tau_1 - t_1) + \tfrac{2}{\theta_1} P_{s,112}(t_1) P_{12,j}(\tau_1 - t_1) + \tfrac{2}{\theta_2} P_{s,221}(t_1) P_{12,j}(\tau_1 - t_1) + 3\tfrac{2}{\theta_2} P_{s,222}(t_1) P_{22,j}(\tau_1 - t_1) \Big], \\
f(G_3, t_0, t_1) &= \tfrac{2}{\theta_4} \mathrm{e}^{-\frac{2}{\theta_5}(\tau_0 - \tau_1)} \mathrm{e}^{-\frac{2}{\theta_4} t_0} \times \\
&\sum_{j \in S_2} \Big[3\tfrac{2}{\theta_1} P_{s,111}(t_1) P_{11,j}(\tau_1 - t_1) + \tfrac{2}{\theta_1} P_{s,112}(t_1) P_{12,j}(\tau_1 - t_1) + \tfrac{2}{\theta_2} P_{s,122}(t_1) P_{12,j}(\tau_1 - t_1) + 3\tfrac{2}{\theta_2} P_{s,222}(t_1) P_{22,j}(\tau_1 - t_1) \Big], \\
f(G_4, t_0, t_1) &= \tfrac{6}{\theta_5} \mathrm{e}^{-\frac{6}{\theta_5} t_1} \, \tfrac{2}{\theta_5} \mathrm{e}^{-\frac{2}{\theta_5} t_0} \times \sum_{j \in S_3} P_{s,j}(\tau_1), \qquad 0 < t_1 + t_0 < \tau_0 - \tau_1, \\
f(G_5, t_0, t_1) &= \tfrac{6}{\theta_5} \mathrm{e}^{-\frac{6}{\theta_5} t_1} \mathrm{e}^{-\frac{2}{\theta_5}(\tau_0 - \tau_1 - t_1)} \, \tfrac{2}{\theta_4} \mathrm{e}^{-\frac{2}{\theta_4} t_0} \times \sum_{j \in S_3} P_{s,j}(\tau_1), \qquad 0 < t_1 < \tau_0 - \tau_1,\ 0 < t_0 < \infty, \\
f(G_6, t_0, t_1) &= \mathrm{e}^{-\frac{6}{\theta_5}(\tau_0 - \tau_1)} \, \tfrac{6}{\theta_4} \mathrm{e}^{-\frac{6}{\theta_4} t_1} \, \tfrac{2}{\theta_4} \mathrm{e}^{-\frac{2}{\theta_4} t_0} \times \sum_{j \in S_3} P_{s,j}(\tau_1), \qquad 0 < t_0, t_1 < \infty,
\end{aligned} \tag{12}$$

where S_2 and S_3 are the sets of states with two and three sequences, respectively, that can be reached by the initial state (Table 2). Again each density for a tree shape should be divided by 3 to give the density for the gene tree: e.g., $f(G_{2a}, t_0, t_1) = f(G_2, t_0, t_1)/3$.

Case II: Initial states 112 and 122

For initial state s = 112 or 122, the likelihood calculation at each locus averages over all 18 gene trees (Table S1). This is the only case in this study where it is necessary to keep track of both the sequence IDs and the population IDs in our Markov chain characterization of the process of coalescent with migration. The initial states are thus $1_a1_b2_c$ or $1_a2_b2_c$. However, for states of three sequences, we always arrange the sequence IDs in the order a, b, and c to simplify the notation and thus the subscripts are dropped. Thus $1_a1_b1_c$, $1_a1_b2_c$ and $1_a2_b2_c$ are written as 111, 112 and 122, respectively. There are 21 states in the chain: 111, 112, 121, 122, 211, 212, 221, 222, $1_{bc}1_a$, $1_{ca}1_b$, $1_{ab}1_c$, $1_{bc}2_a$, $1_{ca}2_b$, $1_{ab}2_c$, 1_a2_{bc}, 1_b2_{ca}, 1_c2_{ab}, $2_{bc}2_a$, $2_{ca}2_b$, $2_{ab}2_c$, and 1|2. The states of two sequences have the subscripts to indicate the sequence IDs. For example, $1_{bc}2_a$ means that sequences *b* and *c* have coalesced and their ancestor is in population 1 while sequence *a* is in population 2.

For gene tree G_{1c}, with $0 < t_0 + t_1 < \tau_1$, we have

$$\begin{aligned}
&f(G_{1c}, t_0, t_1)\\
&= \tfrac{2}{\theta_1} P_{s,111}(t_1)\Big(\tfrac{2}{\theta_1} P_{1_{ab}1_c,1_{ab}1_c}(t_0) + \tfrac{2}{\theta_2} P_{1_{ab}1_c,2_{ab}2_c}(t_0)\Big) + \tfrac{2}{\theta_1} P_{s,112}(t_1)\Big(\tfrac{2}{\theta_1} P_{1_{ab}2_c,1_{ab}1_c}(t_0) + \tfrac{2}{\theta_2} P_{1_{ab}2_c,2_{ab}2_c}(t_0)\Big)\\
&+ \tfrac{2}{\theta_2} P_{s,221}(t_1)\Big(\tfrac{2}{\theta_1} P_{1_c2_{ab},1_{ab}1_c}(t_0) + \tfrac{2}{\theta_2} P_{1_c2_{ab},2_{ab}2_c}(t_0)\Big) + \tfrac{2}{\theta_2} P_{s,222}(t_1)\Big(\tfrac{2}{\theta_1} P_{2_{ab}2_c,1_{ab}1_c}(t_0) + \tfrac{2}{\theta_2} P_{2_{ab}2_c,2_{ab}2_c}(t_0)\Big),
\end{aligned} \qquad (13)$$

The densities for gene trees G_{1b} and G_{1a} are similar.

$$\begin{aligned}
&f(G_{1b}, t_0, t_1)\\
&= \tfrac{2}{\theta_1} P_{s,111}(t_1)\Big(\tfrac{2}{\theta_1} P_{1_{ca}1_b,1_{ca}1_b}(t_0) + \tfrac{2}{\theta_2} P_{1_{ca}1_b,2_{ca}2_b}(t_0)\Big) + \tfrac{2}{\theta_2} P_{s,212}(t_1)\Big(\tfrac{2}{\theta_1} P_{1_b2_{ca},1_{ca}1_b}(t_0) + \tfrac{2}{\theta_2} P_{1_b2_{ca},2_{ca}2_b}(t_0)\Big)\\
&+ \tfrac{2}{\theta_1} P_{s,121}(t_1)\Big(\tfrac{2}{\theta_1} P_{1_{ca}2_b,1_{ca}1_b}(t_0) + \tfrac{2}{\theta_2} P_{1_{ca}2_b,2_{ca}2_b}(t_0)\Big) + \tfrac{2}{\theta_2} P_{s,222}(t_1)\Big(\tfrac{2}{\theta_1} P_{2_{ca}2_b,1_{ca}1_b}(t_0) + \tfrac{2}{\theta_2} P_{2_{ca}2_b,2_{ca}2_b}(t_0)\Big),
\end{aligned} \qquad (14)$$

$$\begin{aligned}
&f(G_{1a}, t_0, t_1)\\
&= \tfrac{2}{\theta_1} P_{s,111}(t_1)\Big(\tfrac{2}{\theta_1} P_{1_{bc}1_a,1_{bc}1_a}(t_0) + \tfrac{2}{\theta_2} P_{1_{bc}1_a,2_{bc}2_a}(t_0)\Big) + \tfrac{2}{\theta_2} P_{s,122}(t_1)\Big(\tfrac{2}{\theta_1} P_{1_a2_{bc},1_{bc}1_a}(t_0) + \tfrac{2}{\theta_2} P_{1_a2_{bc},2_{bc}2_a}(t_0)\Big)\\
&+ \tfrac{2}{\theta_1} P_{s,211}(t_1)\Big(\tfrac{2}{\theta_1} P_{1_{bc}2_a,1_{bc}1_a}(t_0) + \tfrac{2}{\theta_2} P_{1_{bc}2_a,2_{bc}2_a}(t_0)\Big) + \tfrac{2}{\theta_2} P_{s,222}(t_1)\Big(\tfrac{2}{\theta_1} P_{2_{bc}2_a,1_{bc}1_a}(t_{01}) + \tfrac{2}{\theta_2} P_{2_{bc}2_a,2_{bc}2_a}(t_0)\Big),
\end{aligned}$$

For gene tree G_2, we have $t_1 < \tau_1$, $t_0 < \tau_0 - \tau_1$, and

$$\begin{aligned}
&f(G_{2c}, t_0, t_1) = \tfrac{2}{\theta_5} \mathrm{e}^{-\frac{2}{\theta_5} t_0} \times\\
&\sum_{j \in S_2} \Big[\tfrac{2}{\theta_1} P_{s,111}(t_1) P_{1_{ab}1_c,j}(\tau_1 - t_1) + \tfrac{2}{\theta_1} P_{s,112}(t_1) P_{1_{ab}2_c,j}(\tau_1 - t_1) + \tfrac{2}{\theta_2} P_{s,221}(t_1) P_{1_c2_{ab},j}(\tau_1 - t_1) + \tfrac{2}{\theta_2} P_{s,222}(t_1) P_{2_{ab}2_c,j}(\tau_1 - t_1)\Big],\\
&f(G_{2b}, t_0, t_1) = \tfrac{2}{\theta_5} \mathrm{e}^{-\frac{2}{\theta_5} t_0} \times\\
&\sum_{j \in S_2} \Big[\tfrac{2}{\theta_1} P_{s,111}(t_1) P_{1_{ca}1_b,j}(\tau_1 - t_1) + \tfrac{2}{\theta_1} P_{s,121}(t_1) P_{1_{ca}2_b,j}(\tau_1 - t_1) + \tfrac{2}{\theta_2} P_{s,212}(t_1) P_{1_b2_{ca},j}(\tau_1 - t_1) + \tfrac{2}{\theta_2} P_{s,222}(t_1) P_{2_{ca}2_b,j}(\tau_1 - t_1)\Big],\\
&f(G_{2a}, t_0, t_1) = \tfrac{2}{\theta_{12}} \tfrac{2}{\theta_5} \mathrm{e}^{-\frac{2}{\theta_5} t_0} \times\\
&\sum_{j \in S_2} \Big[\tfrac{2}{\theta_1} P_{s,111}(t_1) P_{1_{bc}1_a,j}(\tau_1 - t_1) + \tfrac{2}{\theta_1} P_{s,211}(t_1) P_{1_{bc}2_a,j}(\tau_1 - t_1) + \tfrac{2}{\theta_2} P_{s,122}(t_1) P_{1_a2_{bc},j}(\tau_1 - t_1) + \tfrac{2}{\theta_2} P_{s,222}(t_1) P_{2_{bc}2_a,j}(\tau_1 - t_1)\Big].
\end{aligned}$$

(15)

For gene tree G_3, with $t_1 < \tau_1 < \tau_0 < t_0$, we have

$$
\begin{aligned}
&f(G_{3c},t_0,t_1)=\mathrm{e}^{-\frac{2}{\theta_5}(\tau_0-\tau_1)}\,\tfrac{2}{\theta_4}\mathrm{e}^{-\frac{2}{\theta_4}t_0}\times\\
&\sum_{j\in S_2}\Big[\tfrac{2}{\theta_1}P_{s,111}(t_1)P_{1_{ab}1_c,j}(\tau_1-t_1)+\tfrac{2}{\theta_1}P_{s,112}(t_1)P_{1_{ab}2_c,j}(\tau_1-t_1)+\tfrac{2}{\theta_2}P_{s,221}(t_1)P_{1_c2_{ab},j}(\tau_1-t_1)+\tfrac{2}{\theta_2}P_{s,222}(t_1)P_{2_{ab}2_c,j}(\tau_1-t_1)\Big],\\
&f(G_{3b},t_0,t_1)=\tfrac{2}{\theta_{12}}\mathrm{e}^{-\frac{2}{\theta_5}(\tau_0-\tau_1)}\,\tfrac{2}{\theta_4}\mathrm{e}^{-\frac{2}{\theta_4}t_0}\times\\
&\sum_{j\in S_2}\Big[\tfrac{2}{\theta_1}P_{s,111}(t_1)P_{1_{ca}1_b,j}(\tau_1-t_1)+\tfrac{2}{\theta_1}P_{s,121}(t_1)P_{1_{ca}2_b,j}(\tau_1-t_1)+\tfrac{2}{\theta_2}P_{s,212}(t_1)P_{1_b2_{ca},j}(\tau_1-t_1)+\tfrac{2}{\theta_2}P_{s,222}(t_1)P_{2_{ca}2_b,j}(\tau_1-t_1)\Big]\\
&f(G_{3a},t_0,t_1)=\tfrac{2}{\theta_{12}}\mathrm{e}^{-\frac{2}{\theta_5}(\tau_0-\tau_1)}\,\tfrac{2}{\theta_4}\mathrm{e}^{-\frac{2}{\theta_4}t_0}\times\\
&\sum_{j\in S_2}\Big[\tfrac{2}{\theta_1}P_{s,111}(t_1)P_{1_{bc}1_a,j}(\tau_1-t_1)+\tfrac{2}{\theta_1}P_{s,211}(t_1)P_{1_{bc}2_a,j}(\tau_1-t_1)+\tfrac{2}{\theta_2}P_{s,122}(t_1)P_{1_a2_{bc},j}(\tau_1-t_1)+\tfrac{2}{\theta_2}P_{s,222}(t_1)P_{2_{bc}2_a,j}(\tau_1-t_1)\big)\Big].
\end{aligned}
$$

$$(16)$$

For gene trees G_4, G_5, and G_6, the probability density does not depend on the sequence IDs.

$$
\begin{aligned}
f(G_{4k},t_0,t_1)&=\tfrac{2}{\theta_5}\mathrm{e}^{-\frac{6}{\theta_5}t_1}\,\tfrac{2}{\theta_5}\mathrm{e}^{-\frac{2}{\theta_5}t_0}\times\sum_{j\in S_3}P_{s,j}(\tau_1), && 0<t_1+t_0<\tau_0-\tau_1,\\
f(G_{5k},t_0,t_1)&=\tfrac{2}{\theta_5}\mathrm{e}^{-\frac{6}{\theta_5}t_1}\mathrm{e}^{-\frac{2}{\theta_5}(\tau_0-\tau_1-t_1)}\,\tfrac{2}{\theta_4}\mathrm{e}^{-\frac{2}{\theta_4}t_0}\times\sum_{j\in S_3}P_{s,j}(\tau_1), && 0<t_1<\tau_0-\tau_1, 0<t_0<\infty,\\
f(G_{6k},t_0,t_1)&=\mathrm{e}^{-\frac{6}{\theta_5}(\tau_0-\tau_1)}\,\tfrac{2}{\theta_4}\mathrm{e}^{-\frac{6}{\theta_4}t_1}\,\tfrac{2}{\theta_4}\mathrm{e}^{-\frac{2}{\theta_4}t_0}\times\sum_{j\in S_3}P_{s,j}(\tau_1), && 0<t_1,t_0<\infty,
\end{aligned}
\qquad (17)
$$

where $k = c$, a, and b.

Case III: Initial states 113, 123, and 223

For initial state s = 113, 123, or 223, only three gene tree shapes are possible: G_3, G_5, and G_6 (Table S1). For tree shapes G_3 and G_5, the only gene tree possible is G_{3c} or G_{5c}: ((*a*, *b*), *c*), while for the tree shape G_6, the three gene trees G_{6c}: ((*a*, *b*), *c*); G_{6a}: ((*b*, *c*), *a*); and G_{6b}: ((*c*, *a*), *b*) have the same prior density. Thus there is no need to trace the sequence IDs. There are four states in the chain: 113, 123, 223, 13|23, with the rate matrix given as follows.

$$
\begin{array}{c|cccc}
 & 113 & 123 & 223 & 13\,|\,23\\
\hline
113 & -(2w_{21}+c_1) & 2w_{21} & 0 & c_1\\
123 & w_{12} & -(w_{12}+w_{21}) & w_{21} & 0\\
223 & 0 & 2w_{12} & -(2w_{12}+c_2) & c_2\\
13\,|\,23 & 0 & 0 & 0 & 0
\end{array}
\qquad (18)
$$

For tree shapes G_3 and G_5, only one gene tree is possible, so that

$$
\begin{aligned}
f(G_{3c},t_0,t_1)&=\tfrac{2}{\theta_4}\mathrm{e}^{-\frac{2}{\theta_4}t_0}\times\Big[\tfrac{2}{\theta_1}P_{s,113}(t_1)+\tfrac{2}{\theta_2}P_{s,223}(t_1)\Big],\\
f(G_{5c},t_0,t_1)&=\tfrac{2}{\theta_5}\tfrac{2}{\theta_4}\mathrm{e}^{-\frac{2}{\theta_5}t_1}\mathrm{e}^{-\frac{2}{\theta_4}t_0}\times\sum_{j\in S_3}P_{s,j}(\tau_1).
\end{aligned}
\qquad (19)
$$

For tree shape G_6, the three gene trees have the same density.

$$
f(G_{6k},\, t_0,\, t_1)=\tfrac{2}{\theta_4}\mathrm{e}^{-\frac{6}{\theta_4}t_1}\,\tfrac{2}{\theta_4}\mathrm{e}^{-\frac{2}{\theta_4}t_0}\times\sum_{j\in S_3}P_{s,j}(\tau_1)\mathrm{e}^{-\frac{2}{\theta_5}\tau_0}\ , \qquad (20)
$$

where $k = c$, a, and b.

Case IV: Initial states 133, 233, and 333

For initial state $s = 133$, 233, or 333, there is no need to trace the sequence IDs. We first discuss the initial state 333. The genealogical process is the single-population coalescent, with different population size parameters: θ_3 for $t < \tau_0$ or θ_4 for $t > \tau_0$. There is no need to distinguish among G_1, G_2, and G_4, or between G_3 and G_5, so we consider only G_1 and G_3, but with the range of the coalescent times modified accordingly. There are thus three tree shapes: G_1, G_3, and G_6. For each one, we sum over three gene trees. Thus with initial state $s = 333$, we have

$$f(G_k, t_0, t_1) = \begin{cases} \frac{2}{\theta_3}\frac{2}{\theta_3}\mathrm{e}^{-\frac{6}{\theta_3}t_1}\mathrm{e}^{-\frac{2}{\theta_3}t_0}, & 0 < t_1 + t_0 < \tau_0, \text{ for } k = 1c, 1a, 1b, \\ \frac{2}{\theta_3}\frac{2}{\theta_4}\mathrm{e}^{-\frac{6}{\theta_3}t_1}\mathrm{e}^{-\frac{2}{\theta_3}(\tau_0 - t_1)}\mathrm{e}^{-\frac{2}{\theta_4}t_0}, & t_1 < \tau_0, \quad \text{for } k = 3c, 3a, 3b, \\ \frac{2}{\theta_4}\frac{2}{\theta_4}\mathrm{e}^{-\frac{6}{\theta_3}\tau_0}\mathrm{e}^{-\frac{6}{\theta_4}t_1}\mathrm{e}^{-\frac{2}{\theta_4}t_0}, & 0 < t_1, t_0 < \infty, \quad \text{for } k = 6c, 6a, 6b. \end{cases} \tag{21}$$

Similarly, for initial state $s = 133$ or 233, we consider two tree shapes G_3 and G_6.

$$f(G_k, t_0, t_1) = \begin{cases} \frac{2}{\theta_3}\frac{2}{\theta_4}\mathrm{e}^{-\frac{2}{\theta_3}t_1}\mathrm{e}^{-\frac{2}{\theta_4}t_0}, & t_1 < \tau_0, \quad \text{for } k = 3, \\ \frac{2}{\theta_4}\frac{2}{\theta_4}\mathrm{e}^{-\frac{2}{\theta_3}\tau_0}\mathrm{e}^{-\frac{6}{\theta_4}t_1}\mathrm{e}^{-\frac{2}{\theta_4}t_0}, & 0 < t_1, t_0 < \infty, \text{ for } k = 6c, 6a, 6b. \end{cases} \tag{22}$$

FIGURE LEGENDS

FIGURE 1. (**a**) Species tree illustrating parameters in model M2 (gene flow) for three species (1, 2, and 3) and (**b**)-(**g**) possible gene tree shapes for a locus with three sequences (*a*, *b*, and *c*). With certain initial states (data configurations at the locus), we have to keep track of the sequence IDs (*a*, *b*, and *c*) as well as the population IDs, so that each gene tree shape may correspond to three distinct gene trees. For example, with the data configuration (initial state) $1_a2_b3_c$, the tree shape G_6 represents three distinct gene trees: G_{6c}: ((*a*, *b*), *c*); G_{6a}: ((*b*, *c*), *a*); and G_{6b}: ((*c*, *a*), *b*).

FIGURE 2. The three gene trees with branch lengths for three sequences *a*, *b*, and *c*. Branch lengths b_0 and b_1 are simple linear functions of coalescent times t_0 and t_1 in the gene trees of Fig. 1. For example, for the tree G_1 of Fig. 1, $b_0 = t_0$ and $b_1 = t_1$, while for G_2, $b_0 = t_0 + \tau_1 - t_1$ and $b_1 = t_1$.

FIGURE 3. Posterior probabilities of the six possible gene trees (G_{3c}, G_{5c}, G_{6c}, G_{6a}, and G_{6b}) for the '123' loci in a dataset simulated using the MLEs of parameters for the *Drosophila* dataset D1 (auto).

FIGURE 4. Posterior probabilities of gene trees for the MSY loci for dataset D5 (exons split). The red lines for gene tree G_{3c} indicated loci that are likely to have been transferred across species, with $P(G_{3c}) > 95\%$.

TABLE 1. Summary of the density for coalescent time for two sequences under M0 (no gene flow)

State	$f(t)$ before transform	t limits	$f(x)$ after transform	x limits	b
11	$\frac{2}{\theta_1}\mathrm{e}^{-\frac{2}{\theta_1}t}$	$(0,\ \tau_1)$	e^{-x}	$(0, \frac{2}{\theta_1}\tau_1)$	$\frac{\theta_1}{2}x$
	$\mathrm{e}^{-\frac{2}{\theta_1}\tau_1}\frac{2}{\theta_5}\mathrm{e}^{-\frac{2}{\theta_5}(t-\tau_1)}$	$(\tau_1,\ \tau_0)$	$\mathrm{e}^{-\frac{2}{\theta_1}\tau_1}\mathrm{e}^{-x}$	$(0, \frac{2}{\theta_5}(\tau_0 - \tau_1))$	$\tau_1 + \frac{\theta_5}{2}x$
	$\mathrm{e}^{-\frac{2}{\theta_1}\tau_1}\mathrm{e}^{-\frac{2}{\theta_5}(\tau_0-\tau_1)}\frac{2}{\theta_4}\mathrm{e}^{-\frac{2}{\theta_4}(t-\tau_0)}$	$(\tau_0,\ \infty)$	$\mathrm{e}^{-\frac{2}{\theta_1}\tau_1}\mathrm{e}^{-\frac{2}{\theta_5}(\tau_0-\tau_1)}\mathrm{e}^{-x}$	$(0, \infty)$	$\tau_0 + \frac{\theta_4}{2}x$
22	As for 11 above, with θ_1 replaced by θ_2				
12	$\frac{2}{\theta_5}\mathrm{e}^{-\frac{2}{\theta_5}(t-\tau_1)}$	$(\tau_1,\ \tau_0)$	e^{-x}	$(0, \frac{2}{\theta_5}(\tau_0 - \tau_1))$	$\tau_1 + \frac{\theta_5}{2}x$
	$\mathrm{e}^{-\frac{2}{\theta_5}(\tau_0-\tau_1)}\frac{2}{\theta_4}\mathrm{e}^{-\frac{2}{\theta_4}(t-\tau_0)}$	$(\tau_0,\ \infty)$	$\mathrm{e}^{-\frac{2}{\theta_5}(\tau_0-\tau_1)}\mathrm{e}^{-x}$	$(0, \infty)$	$\tau_0 + \frac{\theta_4}{2}x$
13/23	$\frac{2}{\theta_4}\mathrm{e}^{-\frac{2}{\theta_4}(t-\tau_0)}$	$(\tau_0,\ \infty)$	e^{-x}	$(0, \infty)$	$\tau_0 + \frac{\theta_4}{2}x$
33	$\frac{2}{\theta_3}\mathrm{e}^{-\frac{2}{\theta_3}t}$	$(0,\ \tau_0)$	e^{-x}	$(0, \frac{2}{\theta_3}\tau_0)$	$\frac{\theta_3}{2}x$
	$\frac{2}{\theta_4}\mathrm{e}^{-\frac{2}{\theta_3}\tau_0}\mathrm{e}^{-\frac{2}{\theta_4}(t-\tau_0)}$	$(\tau_0,\ \infty)$	$\mathrm{e}^{-x}\mathrm{e}^{-\frac{2}{\theta_3}\tau_0}$	$(0, \infty)$	$\tau_0 + \frac{\theta_4}{2}x$

TABLE 2. **Markov chains and their states for characterizing the genealogical process of epoch E_1 in model M2 (gene flow)**

Case	Initial states	States in chain	Calculation of $P(t)$
	Loci with 3 sequences		
I	{111, 222}	{111, 112, 122, 222, 11, 12, 22, 1|2} 8 states	Numerical
II	{112, 122}	{111, 112, 121, 122, 211, 212, 221, 222, $1_{bc}1_a$, $1_{ca}1_b$, $1_{ab}1_c$, $1_{bc}2_a$, $1_{ca}2_b$, $1_{ab}2_c$, 1_a2_{bc}, 1_b2_{ca}, 1_c2_{ab}, $2_{bc}2_a$, $2_{ca}2_b$, $2_{ab}2_c$, 1|2} 21 states	Numerical
III	{113, 123, 223}	{113, 123, 223, 13|23}	Numerical
IV	{133, 233, 333}	{133, 233, 13, 23, 33, 3}	Analytical
	Loci with 2 sequences		
V	{11, 12, 22}	{11, 12, 22, 1|2}	Numerical
VI	{13, 23, 33}	{13, 23, 33, 3}	Analytical

Note.— In case II (with initial states 112 or 122), it is necessary to keep track of both the population ID (1, 2, 3) and the sequence ID (*a*, *b*, *c*), so that state $1_{ab}2_c$ means two lineages in the sample, with the common ancestor of *a* and *b* in population 1, and sequence *c* in population 2.

TABLE 3. Rate matrix *Q* for the Markov chain for initial states 111 and 222 under model M2

	111	112	122	222	11	12	22	1|2
111	·	$3 \times 4M_{21}/\theta_1$			$3 \times 2/\theta_1$			
112	$4M_{12}/\theta_2$	·	$2 \times 4M_{21}/\theta_1$			$2/\theta_2$		
122		$2 \times 4M_{12}/\theta_2$	·	$4M_{21}/\theta_1$		$2/\theta_1$		
222			$3 \times 4M_{12}/\theta_2$	·			$3 \times 2/\theta_2$	
11					·	$2 \times 4M_{21}/\theta_1$		$2/\theta_1$
12					$4M_{12}/\theta_2$	·	$4M_{21}/\theta_1$	
22						$2 \times 4M_{12}/\theta_2$	·	$2/\theta_2$
1|2								·

Note.— We define parameters using the real-world process (with time running forward), so that the migration rate $M_{ij} = N_j m_{ij}$ is the expected number of migrant individuals from populations *i* to *j* per generation (in the real world) and m_{ij} is the proportion of individuals in population *j* that are immigrants from population *i*. The Markov chain is then used to describe the process of coalescent with migration, with time running backwards. For example $Q_{111,\,112}$ is the rate for the transition from state 111 to state 112, which in the real world means one of the three sequences in population 1 is an immigrant from population 2, which has the rate $3m_{21}$ per generation. Since time is measured by the mutational distance and one time unit is the expected time to accumulate one mutation per site (that is, one time unit is $1/\mu$ generations), the rate per time unit is $Q_{111,\,112} = 3m_{21} \times 1/\mu = 3 \times 4N_1m_{21}/(4N_1\mu)$ $=3 \times 4M_{21}/\theta_1$, as in the table. Given the rate matrix $Q = \{Q_{ij}\}$, the transition probability matrix over time *t* is given as $P(t) = \{P_{ij}(t)\} = \mathrm{e}^{Qt}$. This is the same calculation as in the Markov chain models for nucleotide substitution such as Jukes and Cantor (Jukes and Cantor, 1969).

TABLE 4. Five *Drosophila* datasets analyzed in this paper

Dataset	#MMY loci	#MSY loci	#SSY loci	Total
D1 auto	378	19,224	9,425	29,027
D2 noncoding	378	14,498	7,211	22,087
D3 chrX	0	4,381	2,105	6,486
D4 exons complete	378	27,200	13,500	41,078
D5 exons split	378	10,979	5,342	16,699

TABLE 5. False positive rate, percentage of zeros, and 95% quantile of the null distribution of the LRT statistic ($2\Delta\ell$) comparing the symmetrical versions of models M0 (no gene flow) and M2 (gene flow)

Data	$L = 10$	100	1000	15,000
Set 1 (hominoid): $\theta_4 = \theta_5 = \theta_{12} = 0.005$, $\tau_0 = 0.006$, $\tau_1 = 0.004$				
(a) 123	0.000 0.829 0.034	0.001 0.641 2.217	0.005 0.528 2.708	0.004 0.506 2.443
(b) 11&12	0.003 0.851 0.578	0.019 0.680 1.528	0.045 0.504 2.542	0.084 0.479 3.492
(c) 113&123	0.002 0.848 0.307	0.027 0.674 2.073	0.037 0.576 2.161	0.035 0.507 2.329
Set 2 (mangroves): $\theta_4 = \theta_5 = \theta_{12} = 0.01$, $\tau_0 = 0.02$, $\tau_1 = 0.01$				
(a) 123	0.001 0.883 0.616	0.006 0.798 1.330	0.009 0.709 2.060	0.004 0.345 1.772
(b) 11&12	0.009 0.881 0.454	0.020 0.741 1.542	0.100 0.439 3.872	0.078 0.570 3.481
(c) 113&123	0.010 0.906 0.418	0.035 0.791 1.983	0.031 0.712 2.013	0.039 0.722 2.136
Set 3: $\theta_4 = \theta_{12} = 0.02$, $\theta_5 = 0.03$, $\tau_0 = 0.06$, $\tau_1 = 0.04$				
(a) 123	0.000 0.957 0.000	0.002 0.904 0.501	0.001 0.896 0.424	0.006 0.884 0.975
(b) 11&12	0.007 0.864 0.796	0.032 0.727 1.979	0.035 0.713 1.814	0.009 0.839 0.422
(c) 113&123	0.003 0.945 0.017	0.008 0.902 0.535	0.007 0.895 0.589	0.008 0.910 0.198
Set 4: $\theta_4 = \theta_{12} = 0.02$, $\theta_5 = 0.01$, $\tau_0 = 0.02$, $\tau_1 = 0.01$				
(a) 123	0.000 0.854 1.137	0.003 0.782 1.469	0.001 0.717 0.841	0.002 0.685 2.003
(b) 11&12	0.008 0.823 0.479	0.032 0.757 1.707	0.047 0.625 2.470	0.049 0.656 2.687
(c) 113&123	0.013 0.823 1.056	0.040 0.775 2.069	0.034 0.719 1.782	0.030 0.666 2.136

Note.— In each cell, the three numbers are the false positive rate, the proportion of replicates in which the test statistic is $2\Delta\ell = 0$, and the estimated 95% critical value. The critical value used for the test is $\chi^2_{2,5\%}$ = 5.99 for (a) configuration 123, and is 2.71 for (b) 11&12 and (c) 113&123.

TABLE 6. Power of the LRT comparing the symmetrical versions of models M0 (no gene flow) and M2 (gene flow)

Data	$L = 10$	100	1000	15,000
Set 1 (hominoid): $\theta_4 = \theta_5 = \theta_{12} = 0.005$, $\tau_0 = 0.006$, $\tau_1 = 0.004$, $M = 1$				
(a) 123	0.6%	5.3%	81.6%	100%
(b) 11&12	4.6%	7.0%	16.1%	65.7%
(c) 113&123	3.3 %	17.9%	88.3%	100%
Set 2 (mangroves): $\theta_4 = \theta_5 = \theta_{12} = 0.01$, $\tau_0 = 0.02$, $\tau_1 = 0.01$, $M = 1$				
(a) 123	3.0%	52.1%	100%	100%
(b) 11&12	8.0%	27.3%	32.0%	89.3%
(c) 113&123	13.8%	69.3%	100%	100%

Note.— The critical value used is 5.99 for (a) 123, and is 2.71 for (b) 11&12 and (c) 113&123.

Table 7. Means and SDs of MLEs from datasets simulated under the symmetrical model M2 (gene flow)

Data	(a) 11&12						(b) 113&123					
	θ_4	θ_5	τ_0	τ_1	θ_{12}	M	θ_4	θ_5	τ_0	τ_1	θ_{12}	M
Set 1 (hominoid): $\theta_4 = \theta_5 = \theta_{12} = 0.005$, $\tau_0 = 0.006$, $\tau_1 = 0.004$, $M = 1$												
Truth	5	5	6	4	5	1	5	5	6	4	5	1
L = 100	6.7 ± 4.1	33.7 ± 191.0	6.7 ± 3.0	3.4 ± 2.3	9.3 ± 64.0	1.4 ± 1.7	4.9 ± 1.0	10.8 ± 90.2	6.0 ± 0.4	3.6 ± 1.9	6.6 ± 8.1	1.3 ± 1.4
L = 1000	5.5 ± 2.5	20.0 ± 152.5	7.4 ± 3.6	3.4 ± 1.9	6.9 ± 56.9	1.1 ± 0.7	5.0 ± 0.3	4.7 ± 2.0	6.0 ± 0.1	4.0 ± 1.2	5.1 ± 0.6	1.1 ± 0.6
L = 15000	5.1 ± 1.0	14.1 ± 98.3	7.4 ± 4.1	3.5 ± 1.3	5.0 ± 0.4	0.9 ± 0.2	5.0 ± 0.1	5.0 ± 0.6	6.0 ± 0.0	4.0 ± 0.3	5.0 ± 0.1	1.0 ± 0.1
Set 2 (mangroves): $\theta_4 = \theta_5 = \theta_{12} = 0.01$, $\tau_0 = 0.02$, $\tau_1 = 0.01$, $M = 1$												
Truth	10	10	20	10	10	1	10	10	20	10	10	1
L = 100	13.1 ± 7.5	17.8 ± 87.2	18.6 ± 7.5	8.8 ± 5.0	10.9 ± 7.3	1.5 ± 1.7	9.9 ± 1.9	9.6 ± 3.9	20.1 ± 0.9	9.9 ± 4.2	14.0 ± 70.0	1.4 ± 1.4
L = 1000	10.9 ± 4.3	13.4 ± 64.5	18.6 ± 7.7	8.6 ± 4.0	10.0 ± 1.7	1.1 ± 0.5	10.0 ± 0.6	9.9 ± 1.2	20.0 ± 0.3	10.0 ± 0.2	10.1 ± 0.6	1.1 ± 0.4
L = 15000	10.1 ± 2.2	16.9 ± 103.4	20.8 ± 7.8	9.5 ± 2.0	10.0 ± 0.2	1.0 ± 0.2	10.0 ± 0.2	10.0 ± 0.3	20.0 ± 0.1	10.0 ± 0.3	10.0 ± 0.1	1.0 ± 0.1

Note.— Estimates of θs and τs are multiplied by 1000. For L = 100 or 1000, some estimates are very large (∞) in certain datasets, causing the mean and SD to be very large. See table 5 for the power of the LRT from the same data.

TABLE **8**. Means and SDs of MLEs from datasets simulated under the asymmetrical IM model M2 (gene flow)

	Parameters (true values in parentheses)							
Data	θ_4 (10)	θ_5 (10)	τ_0 (20)	τ_1 (10)	θ_1 (5)	θ_2 (10)	M_{12} (0.1)	M_{21} (1)
				(a) 223&123				
L = 100	9.9 ± 2.0	16.8 ± 63.1	20.1 ± 0.9	10.4 ± 5.0	9.7 ± 19.3	9.4 ± 5.9	0.2 ± 0.5	1.2 ± 0.8
L = 1000	10.0 ± 0.6	12.6 ± 38.9	20.0 ± 0.3	10.0 ± 4.9	9.5 ± 22.0	9.6 ± 1.6	0.2 ± 0.2	1.6 ± 2.6
L = 15000	10.0 ± 0.2	9.7 ± 1.2	20.0 ± 0.1	10.3 ± 2.9	5.4 ± 3.5	10.0 ± 0.4	0.1 ± 0.0	1.1 ± 0.7
				(b) 113&223&123				
L = 99	9.8 ± 2.0	10.9 ± 26.9	20.1 ± 1.0	10.2 ± 5.0	7.5 ± 5.8	9.3 ± 6.1	0.4 ± 1.0	1.4 ± 1.5
L = 999	10.0 ± 0.6	11.8 ± 37.6	20.0 ± 0.3	10.1 ± 4.7	5.4 ± 1.3	9.5 ± 2.1	0.2 ± 0.2	1.0 ± 0.3
L = 15000	10.0 ± 0.1	9.7 ± 1.3	20.0 ± 0.1	10.1 ± 2.8	5.0 ± 0.3	9.9 ± 0.5	0.1 ± 0.1	1.0 ± 0.1

Note.— Estimates of θs and τs are multiplied by 1000. For $L \leq 1000$, several datasets produced large estimates of θ_5 at the upper bound set by the program. The means and SDs were calculated by excluding those estimates.

TABLE 9. **MLEs and standard errors from the five *Drosophila* datasets of Table 4**

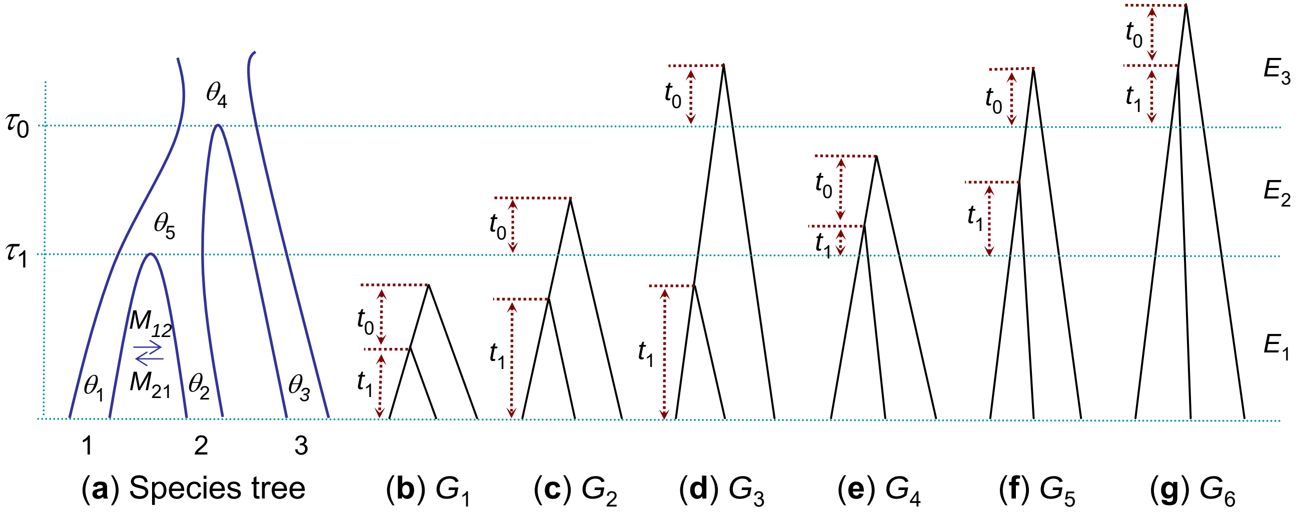
Data & model	τ_{MSY}	τ_{MS}	θ_{MSY}	θ_{MS}	θ_M	θ_S	M_{MS}	M_{SM}	ℓ	$2\Delta\ell$
D1 auto										
M0	24.6 ± 0.1	11.3 ± 0.1	39.4 ± 0.3	13.3 ± 0.2	6.0 ± 0.4	12.8 ± 0.2			–4,763,806.0	
M2	24.3 ± 0.1	13.6 ± 0.2	40.0 ± 0.3	10.6 ± 0.3	5.2 ± 0.6	12.7 ± 0.2	0.0	18.3 ± 3.1	–4,763,452.5	707.0
D2 noncoding										
M0	24.5 ± 0.1	10.8 ± 0.1	41.6 ± 0.4	13.9 ± 0.2	6.0 ± 0.4	13.1 ± 0.2			–3,326,330.8	
M2	24.3 ± 0.1	12.6 ± 0.2	42.1 ± 0.4	12.0 ± 0.2	5.3 ± 0.4	13.0 ± 0.2	0.0	16.2 ± 2.5	–3,326,145.1	371.2
D3 chrX										
M0	28.0 ± 0.2	12.3 ± 0.2	41.1 ± 0.6	15.3 ± 0.4	NA	8.2 ± 0.2			–1,027,233.4	
M2	27.8 ± 0.2	14.2 ± 0.3	41.6 ± 0.6	13.0 ± 0.5	20.9 ± 9.4	8.3 ± 0.2	0.0	40.2 ± 16.9	–1,027,161.6	143.5
M2 ($\theta_M = \theta_S/2$)	27.8 ± 0.2	14.2 ± 0.3	41.6 ± 0.6	13.0 ± 0.5	4.1 ± NA	8.3 ± 0.2	0.0	8.0 ± 1.1	–1,027,161.7	143.5
M2 ($\theta_M = \theta_S$)	27.8 ± 0.2	14.2 ± 0.3	41.6 ± 0.6	13.0 ± 0.5	8.3 ± 0.2		0.0	15.9 ± NA	–1,027,161.7	143.5
D4 exons complete										
M0	20.2 ± 0.1	10.9 ± 0.1	33.7 ± 0.2	9.9 ± 0.1	5.9 ± 0.4	10.7 ± 0.1			–7,853,901.6	
M2	18.3 ± 0.1	18.3 ± 0.1	38.2 ± 0.2	0.0 ± 0.0	4.5 ± 0.5	10.7 ± 0.1	0.0	43.6 ± 4.0	–7,853,313.7	1175.8
M2 ($\tau_{MSY} = 0.020$, $\tau_{MS} = 0.013$)	20	13	34.3 ± 0.2	7.4 ± 0.0	5.1 ± NA	10.6 ± 0.1	0.0	20.7 ± NA	–7,853,425.1	952.9
D5 exons split (subset of D4)										
M0	19.6 ± 0.1	10.9 ± 0.1	38.9 ± 0.3	9.4 ± 0.2	5.9 ± 0.4	10.2 ± 0.2			–2,139,639.5	
M2	18.0 ± 0.1	18.0 ± 0.1	42.6 ± 0.4	0.0 ± 0.0	4.2 ± 0.3	10.2 ± 0.2	0.0	37.8 ± 2.9	–2,139,182.0	915.1
M2 ($\tau_{MSY} = 0.020$, $\tau_{MS} = 0.013$)	20	13	38.5 ± 0.3	7.4 ± 0.4	4.7 ± 0.4	10.1 ± 0.2	0.0	20.4 ± 3.3	–2,139,414.4	450.2

Note.— Estimates of τ, θ, and M are multiplied by 1000. See Table 4 for information about the datasets.

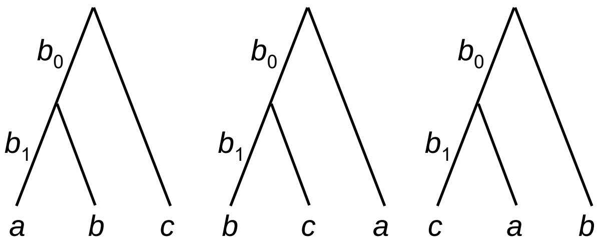
TABLE 10. **MLEs and log likelihood values under M2 assuming different species trees for dataset D1 (auto) of Table 4**

Species tree	τ_{MSY}	τ_1	θ_{MSY}	θ_5	θ_M	θ_S	θ_Y	M_{12}	M_{21}	ℓ
((MS)Y)	24.3 ± 0.1	13.6 ± 0.2 (τ_{MS})	40.0 ± 0.3	10.6 ± 0.3 (θ_{MS})	5.2 ± 0.6	12.7 ± 0.2	NA	0.0 (M_{MS})	18.3 ± 3.1 (M_{SM})	–4,763,452.5
((MY)S)	10.7 ± 0.1	10.7 ± 1.0 (τ_{MY})	53.5 ± 0.3	∞ (θ_{MY})	5.7 ± 0.4	∞	8.2 ± 0.1	0.0 (M_{MY})	0.0 (M_{YM})	–4,780,884.0
((SY)M)	11.4 ± 0.1	11.4 ± 0.1 (τ_{SY})	52.8 ± 0.3	∞ (θ_{SY})	11.3 ± 0.1	∞	4.2 ± 0.3	0.0 (M_{SY})	0.0 (M_{YS})	–4,783,156.2

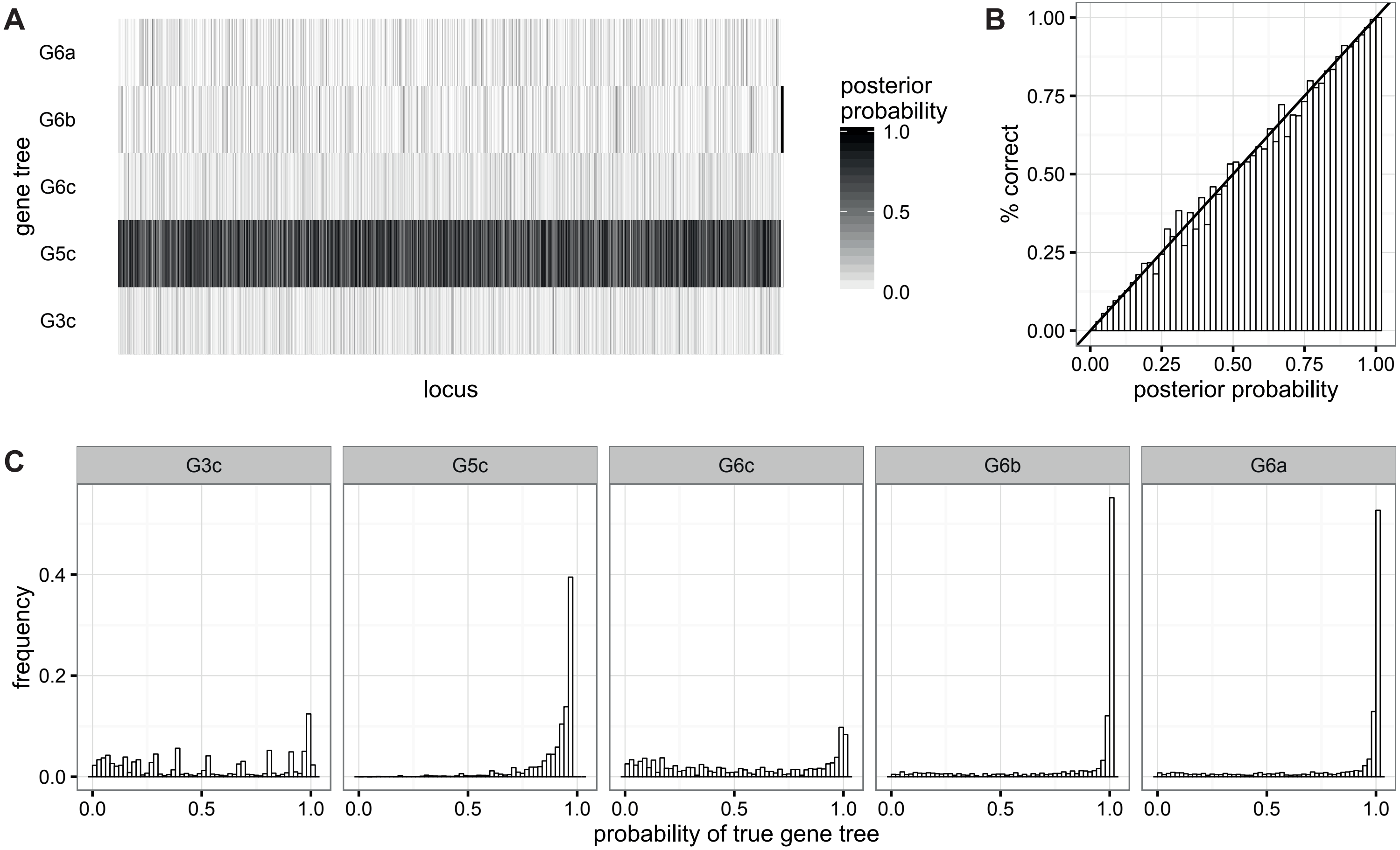
Note.— Estimates of τ, θ, and M are multiplied by 1000. Estimates of θ_5 and θ_S hit the upper bound set in the program for trees ((MY)S) and ((SY)M).

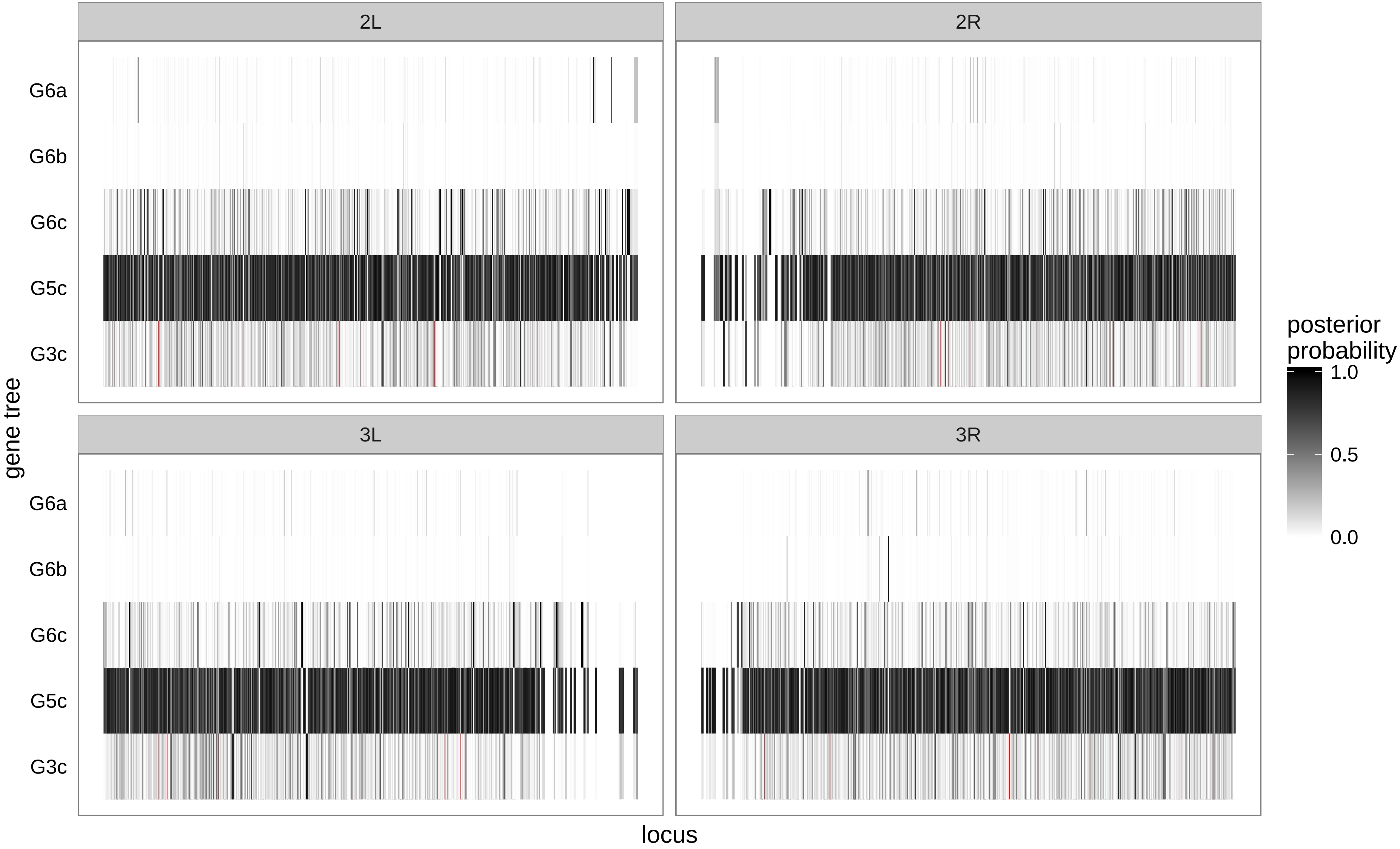


(**a**) Species tree (**b**) G_1 (**c**) G_2 (**d**) G_3 (**e**) G_4 (**f**) G_5 (**g**) G_6



(**a**) T_c (**b**) T_a (**c**) T_b


A
G6a
G6b
G6c
G5c
G3c
gene tree
locus
posterior probability
1.0
0.5
0.0
B
% correct
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
posterior probability
C
G3c
G5c
G6c
G6b
G6a
frequency
0.4
0.2
0.0
0.0
0.5
1.0
probability of true gene tree


2L
2R
3L
3R
G6a
G6b
G6c
G5c
G3c
gene tree
locus
posterior probability
1.0
0.5
0.0