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

Daniel A. Dalquen1, Tianqi Zhu2, and Ziheng Yang1,2,*

1Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK;
2Center for Computational Genomics, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China;
*Correspondence to be sent to: Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London WC1E 6BT, England. E-mail: z.yang@ucl.ac.uk.

Received 21 September 2015; reviews returned 12 January 2016; accepted 8 July 2016

Associate Editor: Thomas Buckley

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 out-group. A Markov chain characterization of the genealogical process of coalescence and migration is used to integrate out the migration histories at each locus analytically, whereas Gaussian quadrature is used to integrate over the coalescent times on each genealogical tree numerically. This is an extension of our earlier 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 data sets 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 in-group species and of the ML estimates of model parameters such as the migration rate. Inclusion of data from a third out-group species is found to increase dramatically the power of the test and the precision of parameter estimation. We compiled and analyzed several genomic data sets 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. [IM model, maximum likelihood, multispecies coalescent, migration, speciation.]

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), whereas 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; Notch and Griffiths 1993; Wilkinson-Herbots 1998). Models of population structure have been implemented in computer programs such as GENETREE (Bahlai 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 data sets have found widespread conflicts among nuclear gene trees and between the mitochondrial gene tree and the nuclear species tree, for example, in mosquitoes (Fontaine et al. 2013), butterflies (Martin et al. 2015), frogs (Zhou et al. 2012), birds (Ellegren et al. 2012), bears (Melo-Ferreira et al. 2012), bears (Liu et al. 2013; Kutschera et al. 2014), and gibbons (Chen et al. 2013). Hybridization both between sister species and between nonsister 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. Although 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 data sets 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 free recombination (Zhu 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 $\theta_1 = \theta_2$), whereas a third species (species 3) is used as the out-group. 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 article 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 data set 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 out-group to improve the power of the analysis. As nicely discussed by Lohse et al. (2011), the out-group 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 out-group 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 (EB) 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 (ML) estimates of model parameters. We use the genome sequences of Drosophila melanogaster, D. simulans, and...
D. yakuba to construct multi-locus data sets 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 article. 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, $\tau_1$), $E_2$: ($\tau_1$, $\tau_3$), and $E_3$: ($\tau_3$, $\infty$) (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 5.

There are nine parameters in the general IM model for three species, including two species divergence times ($\tau_0$ and $\tau_1$), five effective population sizes ($\theta_1$, $\theta_2$, $\theta_3$, $\theta_4$, $\theta_5$), and two migration rates ($M_{12}$ and $M_{21}$). Here $\tau_0$ and $\tau_1$ are scaled by the mutation rate and are measured by the expected number of mutations per site, and $\theta_i = 4N_i\mu$ ($i = 1, \ldots, 5$) are the population size parameters for the five species, with $N_i$ being the (effective) population size of species $i$ and $\mu$ the mutation rate per site per generation. The migration rate is $M_{ij} = N_{ij}m_{ij}$, where $m_{ij}$ is the proportion of immigrants from 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 $\Theta_{12} = \{\tau_0$, $\tau_1$, $\theta_1$, $\theta_2$, $\theta_3$, $\theta_4$, $\theta_5$, $M_{12}$, $M_{21}\}$. Model 0 (no gene flow) is a special case of M2. With $M_{12} = M_{21} = 0$, with parameters $\Theta = \{\tau_0$, $\tau_1$, $\theta_1$, $\theta_2$, $\theta_3$, $\theta_4$, $\theta_5\}$. Note that the symmetrical versions of M0 and M2 assume $\tau_1 = \theta_2$ 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 labeled 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 $L_{123}$. The Markov chain runs backwards in time, describing the change of states due to coalescent and migration. For example, a locus with initial state $L_{123}$ may enter the state $L_{253}$, 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, whereas sequence $c$ is in population 3. There are six gene tree shapes for three sequences: $G_a$–$G_6$ (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 $G_a$ corresponds to three gene trees: $G_{6a}$: ((a, b), c); $G_{6b}$: ((b, c), a); and $G_{6c}$: ((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 (e.g., for a “123” locus, only gene trees $G_{3a}$, $G_{3b}$, $G_{6a}$, $G_{6b}$ are possible under M2), and furthermore some of the gene trees have the same probability distribution under the model (such as $G_{6a}$, $G_{6b}$, and $G_{6c}$). 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, e.g., through comparison with an out-group 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, yxy, and xyy, 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

\[
\begin{align*}
  f(D_i|\theta) &= \sum_k \int_0^{\tau_0} \int_0^{\tau_1} P(D_i|G_k, b_0, b_1)f(G_k, t_0, t_1|\theta)dt_0dt_1 \\
                  &= \int_0^{\tau_0} \int_0^{\tau_1} P(D_i|G_k, b_0, b_1)f(G_k, t_0, t_1|\theta)dt_0dt_1, \\
\end{align*}
\]

(1)

where the sum is over all possible gene trees for the locus, whereas the integrals are over the coalescent times \( t_0 \) and \( t_1 \), with the integral limits \( t_0 \in (0, t_0) \) and \( t_1 \in (t_0, t_1) \) 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 (M1 or M2) and will be described in the next section. For data configurations with three sequences, there are up to \( 6 \times 3 = 18 \) 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

\[
\ell(\theta; D) = \sum_{i=1}^{L} \log f(D_i|\theta).
\]

(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 M0 (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 involves only four parameters: \( \theta = \{t_0, t_1, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5\} \) Note that the population size parameter for a modern species \( (\theta_1, \theta_2, \theta_3) \) 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

\[
f(G_k, t_0, t_1) = \text{rates} \times e^{-T} = \frac{2 \theta_1^2}{\Delta_1} e^{-T},
\]

(3)

where parameters \( \theta_i \) and \( \theta_j \) 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 \( \theta \) is \( 2/\theta \); for very small \( \Delta_i \), the probability that the pair will coalesce during the time interval \( (t_i + \Delta_i) \) is \( 2/\Delta_i \).

Take, for example, configuration “111,” with the initial state \( 1_{a11} \). The probability of data for the locus (Equation (1)) is an average over \( 6 \times 3 \) gene trees. For example, in the case of gene tree \( G_{111} \), the probability density of the gene tree (with coalescent times) is

\[
f(G_{111}, t_0, t_1) = \frac{2 \theta_1}{\theta_1} e^{-T} = \frac{2 \theta_1}{\theta_1} e^{-\frac{2 \theta_1}{\theta_1}} = \frac{2 \theta_1}{\theta_1},
\]

(4)

where \( \frac{2 \theta_1}{\theta_1} \) and \( \frac{2 \theta_1}{\theta_1} \) 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 \( G_{111}, G_{112}, \) and \( G_{112} \). The densities and rates for all data configurations and gene trees are summarized in Supplementary Table S1 (available at Dryad by http://dx.doi.org/10.5061/dryad.hf4vs in Supplementary Material). Note that some gene trees are not possible for certain configurations of loci (e.g., gene trees \( G_{112}, G_{113}, \) and \( G_{114} \) for “112” loci).

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

\[
\int_{x_0}^{x_1} \int_{t_0}^{t_1} P(D_i|G_k, b_0, b_1)f(G_k, t_0, t_1|\theta)dt_0dt_1 = \int_{x_0}^{x_1} \int_{t_0}^{t_1} P(D_i|G_k, b_0, b_1)f(G_k, x_0, x_1) \left| \frac{d(t_0+t_1)}{dx_0} \right| dx_0dx_1.
\]

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

<table>
<thead>
<tr>
<th>State</th>
<th>$f(t)$ before transform</th>
<th>$t$ limits</th>
<th>$f(x)$ after transform</th>
<th>$x$ limits</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>$2 \theta_1$ $\int_0^t$</td>
<td>$(0, \tau_1)$</td>
<td>$e^{-x}$</td>
<td>$(0, \frac{2 \theta_1}{\tau_1})$</td>
<td>$\frac{\theta_1}{\tau_1}$</td>
</tr>
<tr>
<td>22</td>
<td>$\frac{2}{\theta_1} \int_0^{\tau_1} e^{-\frac{x}{\theta_1}}$</td>
<td>$(t_1, \tau_0)$</td>
<td>$\frac{2}{\theta_1} \int_0^{\tau_1} e^{-\frac{x}{\theta_1}}$</td>
<td>$(0, \frac{2 \theta_1}{\tau_1})$</td>
<td>$\frac{\theta_1}{\tau_1}$</td>
</tr>
<tr>
<td>12</td>
<td>$\frac{2}{\theta_1} \int_0^{\tau_1} e^{-\frac{x}{\theta_1}}$</td>
<td>$(t_1, \tau_0)$</td>
<td>$\frac{2}{\theta_1} \int_0^{\tau_1} e^{-\frac{x}{\theta_1}}$</td>
<td>$(0, \frac{2 \theta_1}{\tau_1})$</td>
<td>$\frac{\theta_1}{\tau_1}$</td>
</tr>
<tr>
<td>13/23</td>
<td>$\frac{2}{\theta_1} \int_0^{\tau_1} e^{-\frac{x}{\theta_1}}$</td>
<td>$(t_1, \tau_0)$</td>
<td>$\frac{2}{\theta_1} \int_0^{\tau_1} e^{-\frac{x}{\theta_1}}$</td>
<td>$(0, \frac{2 \theta_1}{\tau_1})$</td>
<td>$\frac{\theta_1}{\tau_1}$</td>
</tr>
<tr>
<td>33</td>
<td>$\frac{2}{\theta_1} \int_0^{\tau_1} e^{-\frac{x}{\theta_1}}$</td>
<td>$(t_1, \tau_0)$</td>
<td>$\frac{2}{\theta_1} \int_0^{\tau_1} e^{-\frac{x}{\theta_1}}$</td>
<td>$(0, \frac{2 \theta_1}{\tau_1})$</td>
<td>$\frac{\theta_1}{\tau_1}$</td>
</tr>
</tbody>
</table>

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 $f(t)$ (Table 1). For instance, with the initial state 11 (two sequences from species 1), $f(t)$ 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

$$f(d_i|t) = \left( \frac{3}{4} - \frac{3}{4} e^{-3t/\lambda} \right) \left( \frac{1}{4} + \frac{3}{4} e^{-3t/\lambda} \right)^{n_i-d_i}.$$  

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

$$f(d_i|\theta) = \int_0^{\infty} f(t|\theta) f(d_i|t) dt.$$  

Gaussian quadrature is used to calculate the 1-D integral, with the transform $x = \frac{3}{4} t$ (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, $f(G_k)$, is given by Equation (6) and is not the same as for Model M0. This is because model M2 uses a different time scale for the coalescent process, which affects the density of the gene trees. The transformation $t = \frac{3}{2} x$ is used to calculate the probabilities in this case.

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 “11,” we use $x_0 = \frac{t_0}{\theta_1} + t_1$, $x_1 = \frac{t_1}{\theta_1} + x_1$ so that $t_0 = \frac{\theta_1}{\lambda} x_0 (1 - x_1)$, $t_1 = \frac{\theta_1}{\lambda} x_1$, The new limits are $0 < x_0 < \frac{\theta_1}{\lambda} t_1, 0 < x_1 < 1,$ and the Jacobian of the transform is $\frac{\partial(t_0, t_1)}{\partial(x_0, x_1)} = \frac{\theta_1}{\lambda^2} \frac{\partial(t_0)}{\partial(x_0)} \frac{\partial(t_1)}{\partial(x_1)}$. Then

$$\int_0^{\tau_1} \int_0^{\tau_1 - t_0} P(D_i|G_1, b_0, b_1) \times \frac{3}{4} \frac{\theta_1}{\lambda} e^{-\frac{3x_0}{\lambda}} \frac{2}{\theta_1} \int_0^{\tau_1} e^{-\frac{x_1}{\theta_1}} dx_1 dx_0 =$$

$$\int_0^{\tau_1} \int_0^{\tau_1 - t_0} P(D_i|G_1, b_0, b_1) \times x_0 e^{-3x_0} \frac{2}{\theta_1} \int_0^{\tau_1} e^{-\frac{x_1}{\theta_1}} dx_1 dx_0,$$

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

The transforms from $(t_0, t_1)$ to $(x_0, x_1)$ are summarized in Supplementary Table S2 in Supplementary Material. 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.
TABLE 2. Markov chains and their states for characterizing the genealogical process of epoch $E_2$ in model M2 (gene flow)

<table>
<thead>
<tr>
<th>Case</th>
<th>Initial states</th>
<th>States in chain</th>
<th>Calculation of $P(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loci with 3 sequences</td>
<td>I (111, 222)</td>
<td>(111, 112, 122, 222, 11, 12, 22, 1, 2) 8 states</td>
<td>Numerical</td>
</tr>
<tr>
<td>II (112, 222)</td>
<td>(111, 112, 121, 221, 222, 121, 122, 11, 12, 22, 1, 2) 12 states</td>
<td>Numerical</td>
<td></td>
</tr>
<tr>
<td>III (113, 123, 223)</td>
<td>(113, 123, 23, 13, 23) 21 states</td>
<td>Numerical</td>
<td></td>
</tr>
<tr>
<td>IV (13, 23, 33)</td>
<td>(13, 23, 13, 23, 33) 3 states</td>
<td>Analytical</td>
<td></td>
</tr>
<tr>
<td>Loci with 2 sequences</td>
<td>V (11, 22, 1</td>
<td>2)</td>
<td>(11, 12, 22, 1, 2) 4 states</td>
</tr>
<tr>
<td>VI (13, 23, 33)</td>
<td>(13, 23, 33) 2 states</td>
<td>Analytical</td>
<td></td>
</tr>
</tbody>
</table>

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

The Markov chain characterization allows one to calculate the probability density for the gene through topology and coalescent times, $f(G_1, t_0, t_1)$, 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 $G_1$: $(a, b, c)$ 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, 222, 11, 12, 22, 1, 2) (Table 2). Here the state “112” 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 $f(G_1, t_0, t_1) = f(G_{1,0}, t_1) = \frac{1}{2} f(G_{1,0}, t_0, t_1)$, with

$$f(G_1, t_0, t_1) = \frac{3}{2} P_{111}(t_1)\left(\frac{2}{3} P_{11,11}(t_0) + \frac{2}{3} P_{11,12}(t_0)\right) + \frac{2}{3} P_{112}(t_1)\left(\frac{2}{3} P_{12,11}(t_0) + \frac{2}{3} P_{12,12}(t_0)\right) + \frac{2}{3} P_{122}(t_1)\left(\frac{2}{3} P_{22,12}(t_0) + \frac{2}{3} P_{22,22}(t_0)\right).$$

Note that the probability density function here has the interpretation that $f(G_1, t_0, t_1) = \Delta t_{1,2}^{-1}$, for very small $\Delta t_{1,2}$, 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 $(t_1 + t_0, t_1 + t_0 + \Delta t_0)$ (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 $P_{111}(t_1)$; the first coalescent occurs in species 1 during $(t_1, t_1 + \Delta t_1)$, with probability $3 \times \frac{1}{3} \Delta t_1$, 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 or in population 2, with probability $P_{11,11}(t_0) \times \left(\frac{2}{3} \Delta t_0\right)$, or in population 2, with probability $P_{12,12}(t_0) \times \left(\frac{2}{3} \Delta t_0\right)$. 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, 222, 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 Supplementary Tables S3 and S4 in Supplementary Material.

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 $P_{123, 1323}(t_1)$ calculated from the Markov chain of Table 2 (case III) is the probability that sequences 1 and 2 have coalesced by time $t_1$. This then gives the probabilities for the five gene trees for the initial state “123” as $P(G_{1,0}) = P_{123, 1323}(t_1)$, $P(G_{2,0}) = P_{23, 1232}(t_1)$, $P(G_{3,0}) = P_{31, 1323}(t_1)$, and $P(G_{0,0}) = 1 - P(G_{1,0}) - 3P(G_{2,0}) + 3P(G_{3,0})$ (Fig. 1). Here $e^{-2/\Delta t_0}$ 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: \( (0, t_1), (t_1, t_0), \) and \( (t_0, \infty), \) so the density is given as

\[
\begin{align*}
 f(t) &= \begin{cases} 
 \frac{2}{\theta_1} P_{11}(t_1) + \frac{2}{\theta_1} P_{22}(t), & t < t_1, \\
 \sum_{j \in B_2} P_{1j}(t_1) \times \frac{2}{\theta_1} e^{-\frac{2}{\theta_1} (t - t_1)}, & t_1 < t < t_0, \\
 \sum_{j \in B_2} P_{1j}(t_1) \times \frac{2}{\theta_2} e^{-\frac{2}{\theta_2} (t - t_0)} \times \frac{2}{\theta_1} e^{-\frac{2}{\theta_1} (t - t_0)}, & t > t_0.
\end{cases}
\end{align*}
\]

(10)

where \( B_2 = \{11, 12, 22\} \) is the set of states with two sequences. The transition probability \( P_{ij}(t) \) is calculated using a Markov chain with four states 11, 12, 22, and 11/2. 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 \( \chi^2 \) approximation is not valid and the null distribution (i.e., 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: \( \theta_1 (\approx \theta_2) \) and \( M. \) However, for two reasons, the large-sample \( \chi^2 \) 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 \( \theta_1 (\approx \theta_2) \) in model M2 becomes unidentified. As a result of the violations of the regularity conditions for the \( \chi^2 \) 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 \( \theta_1 (\approx \theta_2) \) 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, and 233, without any loci of configurations 11, 22, 111, 121, 211, or 223. All such data sets will show the unidentifiability problem under M2 and the two violations of the regularity conditions for the \( \chi^2 \) asymptotics. In this study, we follow Zhu and Yang (2012) and use \( \chi^2 \) 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 \( \theta_1 (\approx \theta_2) \) is identifiable in both models M0 and M2. Although 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 \( \chi^2, \) 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 in Supplementary Material.

A similar unidentifiability problem exists under the asymmetrical model for certain combinations of loci. Let \( U_1 = \{11, 111, 112, 113\} \) and \( U_2 = \{22, 222, 222, 223\}. \) If a data set 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 \( (M_{12} \text{ and } M_{23}) \) than \( M0 \) and a 50:50 mixture of 0 and \( \chi^2 \) is the null distribution, with the significance value \( 2 \Delta \ell = 4.61. \) If a data set 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 \( U_1 \).
state in $U_2$ but none in $U_1$, the model is unidentifiable. In this case the null distribution is unknown and we use $\chi^2$ to conduct the test, with critical value 7.82. If a data set contains none of the states in either $U_1$ or $U_2$, we use $\chi^2$ to conduct the test, with the critical value 9.49 because $M_0$ and $M_2$ differ by four parameters. The critical values for the likelihood ratio test under the asymmetric model for different mixtures of loci are given in Supplementary Table S6 in Supplementary Material.

**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 EB 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 $G_{6a}$, $G_{6b}$, $G_{6c}$, and $G_{6d}$, with $G_{6a}$ 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.

$$f(G_k|D_i, \theta) = \frac{f(G_k|\theta)f(D_i|G_k, \theta)}{f(D_i|\theta)} = \frac{\int f(D_i|G_k, \theta) \int P(D_i|G_k, \theta) \int f(G_k, \theta | \theta) d\theta_1 d\theta_0}{\int \int \int f(D_i|G_k, \theta) \int P(D_i|G_k, \theta) \int f(G_k, \theta | \theta) d\theta_1 d\theta_0}$$

We replace the parameters $(\theta)$ 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 data set 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. 1997) 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 data sets using the parameter values: $\theta_0 = 0.0243$, $\theta_1 = 0.0136$, $\theta_2 = 0.0400$, $\theta_3 = 0.0106$, $\theta_4 = 0.0052$, $\theta_5 = 0.0127$, $M_{12} = 0$ and $M_{21} = 0.0183$, which are the MLEs under $M_2$ from the *Drosophila* data set D1 (auto), to be described and analyzed later (Tables 4 and 9).

We simulated two replicate data sets, each of the same size and configurations as the real data. The results are very similar between the two data sets so we discuss only those for the first data set. The MLEs from the simulated data set are $\hat{\theta}_0 = 0.0242$, $\hat{\theta}_1 = 0.0137$, $\hat{\theta}_2 = 0.0402$, $\hat{\theta}_3 = 0.0104$, $\hat{\theta}_4 = 0.0058$, $\hat{\theta}_5 = 0.0126$, $M_{12} = 0.0018$, and $M_{21} = 0.0196$, 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. Although 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 $G_{6c}$, $G_{6b}$, and $G_{6d}$. Finally, approximately equal proportions of loci are inferred to have gene trees $G_{6a}$, $G_{6b}$, and $G_{6d}$ (Fig. 3a), and they are also close to the expected proportions. Overall the results indicate a well-behaved method.

### Program Implementation, Validation, and Availability

Although 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 $M_0$ (no gene flow) and $M_2$ (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 Supplementary Tables S1 and S2 in Supplementary Material for $M_0$ and similarly in Supplementary Tables S3 and S4 in Supplementary Material for $M_2$.

We conducted extensive tests to validate our implementation. The MCCOMAL program, which is part of the RPP package (Yang and Rannala 2010; Zhang et al. 2011), was used to simulate sequence data under models $M_0$ and $M_2$ 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 data

---

**Table 4.** Five Drosophila data sets analyzed in this article

<table>
<thead>
<tr>
<th>Data set</th>
<th>No. of MMY loci</th>
<th>No. of MSY loci</th>
<th>No. of SSY loci</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1 auto</td>
<td>378</td>
<td>19,224</td>
<td>9,425</td>
<td>29,027</td>
</tr>
<tr>
<td>D2 noncoding</td>
<td>378</td>
<td>14,498</td>
<td>7,211</td>
<td>22,087</td>
</tr>
<tr>
<td>D3 chrX</td>
<td>0</td>
<td>4,381</td>
<td>2,105</td>
<td>6,486</td>
</tr>
<tr>
<td>D4 exons complete</td>
<td>378</td>
<td>27,200</td>
<td>13,500</td>
<td>41,076</td>
</tr>
<tr>
<td>D5 exons split</td>
<td>378</td>
<td>10,979</td>
<td>5,342</td>
<td>16,699</td>
</tr>
</tbody>
</table>
set (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). Although both optimizations are optional, they offer significant speed-ups on genome-scale data sets (Supplementary Fig. S1 in Supplementary Material). 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 Data sets**

We compiled multi-locus data sets 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 data sets 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 nonoverlapping complete exons on chromosomes 2 and 3. When exons were overlapping, only the longest was kept. For all data sets except D4 (exons complete), sequences were split into chunks between 100 bp and 500 bp that were separated by at least 2 kb. 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 data sets 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 paralogs. 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. Data set D5 (exons complete) into loci of between 100 bp and 500 bp 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 2600 bp, and were split into more than one locus in D5. Finally, we added the 378 MMY loci from Hutter et al. (2007) and were split into more than one locus in D5. Finally, some loci in D4 (374 of them) were longer than 2600 bp, and were split into more than one locus in D5. Finally, we added the 378 MMY loci from Hutter et al. (2007) and were split into more than one locus in D5. Finally, we added the 378 MMY loci from Hutter et al. (2007) and were split into more than one locus in D5.

Note: In each cell, the three numbers are the false positive rate, the proportion of replicates in which the test statistic is 2 for data of configuration 123, and is 2.71 for (b) 11&12 and (c) 113&123.

To examine the false positive rate of the test, we simulated replicate data sets 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 200, 1000, 15,000, with each locus having 500 sites. Gene trees with branch lengths (coalset times) were generated from the multispecies coalescent model (Rannala and Yang 2003) using the program MCOAL, which is part of the MrBayes package (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 data set thus consisted of L sequence alignments, with 500 base pairs at each locus. We considered three kinds of data: (i) all loci of configuration 123, (ii) a mixture of loci of configurations 11 and 12 in equal proportions, and (iii) a mixture of loci of configurations 113 and 123 in equal proportions. The number of replicates was 1000.

Overall, the use of the \( \chi^2_2 \) distribution for data of configuration (i) 123 made the test conservative, as the false positive rate was always < 1%, whereas an error rate of 5% was allowed (Table 5). For the “pairs” data (configuration b, 11&12), we observed false positive rates of up to 10% for parameter sets 2 and 3. The analysis

### 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)

<table>
<thead>
<tr>
<th>Data</th>
<th>L = 10</th>
<th>100</th>
<th>1000</th>
<th>15,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1 (hominoids): θ1 + θ2 = 0.005, τ1 = 0.004</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a) 123</td>
<td>0.001 0.889 0.014</td>
<td>0.001 0.798 0.011</td>
<td>0.001 0.798 0.011</td>
<td>0.001 0.798 0.011</td>
</tr>
<tr>
<td>(b) 11&amp;12</td>
<td>0.001 0.896 0.005</td>
<td>0.001 0.798 0.004</td>
<td>0.001 0.798 0.004</td>
<td>0.001 0.798 0.004</td>
</tr>
<tr>
<td>(c) 113&amp;123</td>
<td>0.001 0.896 0.005</td>
<td>0.001 0.798 0.004</td>
<td>0.001 0.798 0.004</td>
<td>0.001 0.798 0.004</td>
</tr>
<tr>
<td>Set 2 (mangroves): θ1 + θ2 = 0.01, τ1 = 0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a) 123</td>
<td>0.001 0.883 0.016</td>
<td>0.001 0.798 0.013</td>
<td>0.001 0.798 0.013</td>
<td>0.001 0.798 0.013</td>
</tr>
<tr>
<td>(b) 11&amp;12</td>
<td>0.001 0.896 0.005</td>
<td>0.001 0.798 0.004</td>
<td>0.001 0.798 0.004</td>
<td>0.001 0.798 0.004</td>
</tr>
<tr>
<td>(c) 113&amp;123</td>
<td>0.001 0.896 0.005</td>
<td>0.001 0.798 0.004</td>
<td>0.001 0.798 0.004</td>
<td>0.001 0.798 0.004</td>
</tr>
<tr>
<td>Set 3: θ1 + θ2 = 0.02, τ1 = 0.03, τ1 = 0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a) 123</td>
<td>0.001 0.957 0.105</td>
<td>0.001 0.896 0.105</td>
<td>0.001 0.896 0.105</td>
<td>0.001 0.896 0.105</td>
</tr>
<tr>
<td>(b) 11&amp;12</td>
<td>0.001 0.957 0.105</td>
<td>0.001 0.896 0.105</td>
<td>0.001 0.896 0.105</td>
<td>0.001 0.896 0.105</td>
</tr>
<tr>
<td>(c) 113&amp;123</td>
<td>0.001 0.957 0.105</td>
<td>0.001 0.896 0.105</td>
<td>0.001 0.896 0.105</td>
<td>0.001 0.896 0.105</td>
</tr>
<tr>
<td>Set 4: θ1 + θ2 = 0.02, τ1 = 0.02, τ1 = 0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a) 123</td>
<td>0.001 0.854 1.107</td>
<td>0.001 0.782 1.107</td>
<td>0.001 0.782 1.107</td>
<td>0.001 0.782 1.107</td>
</tr>
<tr>
<td>(b) 11&amp;12</td>
<td>0.001 0.854 1.107</td>
<td>0.001 0.782 1.107</td>
<td>0.001 0.782 1.107</td>
<td>0.001 0.782 1.107</td>
</tr>
<tr>
<td>(c) 113&amp;123</td>
<td>0.001 0.854 1.107</td>
<td>0.001 0.782 1.107</td>
<td>0.001 0.782 1.107</td>
<td>0.001 0.782 1.107</td>
</tr>
</tbody>
</table>

Note: In each cell, the three numbers are the false positive rate, the proportion of replicates in which the test statistic is 2 for data of configuration 123, and is 2.71 for (b) 11&12 and (c) 113&123.
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 out-group 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 M2 (gene flow). We used parameter values of Set 1 in the data, reducing the false positive rate to below 5%.

For each of the five data sets (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 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.

**Table 6.** Power of the LRT comparing the symmetrical versions of models M0 (no gene flow) and M2 (gene flow) of (hominoid) and Set 2 (mangroves), with M2 (gene flow). We used parameter values of Set 1 in the data, reducing the false positive rate to below 5%.

<table>
<thead>
<tr>
<th>Data</th>
<th>L = 10</th>
<th>100</th>
<th>1000</th>
<th>15,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1 (hominoid): ( \theta_1 = \theta_3 = 0.005, \tau_1 = 0.006, \rho_1 = 0.004, M = 1 )</td>
<td>(a) 123</td>
<td>0.6%</td>
<td>5.3%</td>
<td>81.6%</td>
</tr>
<tr>
<td>(b) 11&amp;12</td>
<td>4.6%</td>
<td>7.0%</td>
<td>16.1%</td>
<td>65.7%</td>
</tr>
<tr>
<td>(c) 13&amp;14</td>
<td>3.3%</td>
<td>17.9%</td>
<td>88.3%</td>
<td>100%</td>
</tr>
<tr>
<td>Set 2 (mangroves): ( \theta_1 = \theta_3 = 0.01, \tau_1 = 0.02, \rho_1 = 0.01, M = 1 )</td>
<td>(a) 123</td>
<td>3.0%</td>
<td>52.1%</td>
<td>100%</td>
</tr>
<tr>
<td>(b) 11&amp;12</td>
<td>8.0%</td>
<td>27.3%</td>
<td>32.0%</td>
<td>89.3%</td>
</tr>
<tr>
<td>(c) 13&amp;14</td>
<td>13.8%</td>
<td>69.3%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

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

Analysis of Drosophila Genomic Data sets

For each of the five data sets (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 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.

Table 7 lists the means and standard deviations of the MLEs of parameters under model M2 for the same data analyzed in Table 6. Data sets 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 \( \theta_1 \) and \( \theta_2 \) 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 data set 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 \( \theta_2, \theta_1, \) and \( M_{21} \). Even with 1000 loci, we encountered several data sets 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 \( \theta_2 = \theta_1 = 0 \) or \( \infty \)). With 15,000 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 15,000 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.

**Table 7.** Means and standard deviations of the MLEs of parameters under model M2 for the same data analyzed in Table 6. Data sets with "123" loci only suffer from the problem of unidentifiability. Inclusion of the "113" loci allows the model to estimate \( \theta_1 \) and \( \theta_2 \) and the unidentifiability problem disappears, leading to better parameter estimation. Furthermore, the "triplet" data provided much better parameter estimates than the "pair" data.
than twice as large as \( \hat{\theta}_K \) (\( \approx 0.005–0.006 \)), 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 data sets D1 (auto) and D2 (noncoding) we obtained \( \hat{\theta}_S = 0.011 \) and \( \hat{\theta}_K = 0.003–0.014 \) under \( M_0 \) and \( \hat{\theta}_S = 0.012–0.014 \) and \( \hat{\theta}_K = 0.011–0.012 \) under \( M_2 \) (Table 9). The slightly smaller estimates of
and large estimates of \( \theta_M \) 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 \( D. simulans \).

Data set D3 (chrX) for the X chromosome showed very different patterns from the autosomal data sets D1 (auto) and D2 (noncoding), with a smaller estimate of \( \theta_S \), and slightly larger estimates of the other \( \theta \) 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 \( \frac{3}{4} \) that for the autosome, so that \( \theta_S \) for X should be \( \frac{3}{4} \) times as large as \( \theta_S \) for the autosomes, whereas the \( MS \) and \( MSY \) should be identical. The parameter estimates suggested that this simplistic model may not fit the data well. However, the estimates of \( \theta_M \) and \( MSM \) from D3 (chrX) were associated with large sampling errors. Indeed, D3 (chrX) does not include any MMM loci, so that the data contain only very weak information concerning \( MSM \) even though the model is identifiable. The correlation between estimates of \( \theta_M \) and \( MSM \) means that estimation of \( MSM \) may be affected as well. We thus reran M2 under the constraint that \( \theta_S = \frac{1}{2} \theta_M \) or \( \theta_N = \theta_M \), obtaining estimates of \( MSM \) to be 0.016 and 0.018 (Table S9). Thus there was no evidence for a large \( MSM \) for the X than for the autosomes. The large changes to \( \theta_M \) and \( MSM \) caused virtually no change to the log likelihood or to estimates of other parameters, suggesting that the data are uninformative about \( \theta_M \) and \( MSM \) 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 data sets, 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 data sets, estimates of \( \tau_S \) 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 \( \tau_{SM} = \tau_{MS} \) and \( \theta_{MS} \approx 0 \) (Table 9). We were unable to determine the reasons for this behavior. We note that the same behavior was encountered in a few simulated data sets, as mentioned earlier, and that the problem did not occur for data set D1 (auto), which includes both coding and noncoding loci. The estimates of \( \theta_M \) and \( \tau_M \) 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 (D1 and D2). We note that estimates of \( \tau \) and \( \theta \) parameters under \( M0 \) from those exon data sets were similar to the \( M0 \) estimates from D1 (auto) and D2 (noncoding), and that the estimates of \( \tau_{MS} \) were very similar to those from D0 and D2 for the same data set. Thus we reran the M2 analysis of the two exon data sets, with \( \tau_{SY} = 0.020 \) and \( \tau_{MS} = 0.013 \) fixed, to estimate the other parameters. The results appear much more reasonable (Table 9). Both data sets D4 and D5 suggested no migration from \( D. melanogaster \) to \( D. simulans \), but the estimates of \( MSM \) at \( \sim 0.02 \) 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 reanalyzed the data sets under \( M2 \) (gene flow) assuming asymmetrical migration rates (with \( MSM \neq MSY \)) but symmetrical population sizes (\( \theta_M = \theta_S \)) (Supplementary Table S7 in Supplementary Material).

Again the LRT is significant in every data set, and parameter estimates suggested unidirectional migration, with \( MSM = 0 \) in every data set. However, estimates of \( MSM \) were much larger than those of Table 9 in every data set except for D3 (chrX), which has been discussed above. For example, \( MSM = 0.036 \)–0.041 from D1 (auto) and D2 (noncoding) under the constraint \( \theta_M = \theta_S \) (Supplementary Table S7 in Supplementary Material), in comparison with 0.016–0.018 without the constraint (Table 9). We note that, except for \( \theta_M \) and \( MSM \), the parameter estimates were virtually identical with and without the constraint \( \theta_M = \theta_S \) (compare Supplementary Tables S7 and S9 in Supplementary Material). There are far more SSV than MMY loci in those data sets (Table 4), so that the estimates of \( \theta_M = \theta_S \), at 0.012 (Supplementary Table S7 in Supplementary Material Table S7), were now dominated by the \( D. simulans \) polymorphism data, and were too large for \( D. melanogaster \). This has led to overestimates of \( MSM \), apparently because a large \( MSM \) is more compatible with the (unrealistically assumed) large \( \theta_M \). Thus the assumption \( \theta_M = \theta_S \) 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 \( \theta_M = \theta_S \); for example, relaxing the assumption improves the log likelihood by 66–82 units in data sets D1 (auto) and D2 (noncoding). D3 (chrX) does not include any MMY loci. As a result, \( \theta_M \) is unidentifiable under \( M0 \) (so that the log likelihood is the same with and without the constraint \( \theta_M = \theta_S \)), whereas under \( M2 \), \( \theta_M \) is identifiable but very poorly estimated (so that the log likelihoods are distinct but extremely similar with and without the constraint) (Supplementary Tables S9 and S7 in Supplementary Material).

We used Equation (11) to calculate the posterior probabilities for gene trees for the MSY loci in the five data sets (Table 4). Here we discuss the results for D5 (exons split) (Fig. 4), and those for D1 (auto) and D3 (chrX) are presented in Supplementary Figures
An intriguing feature in Fig. 4 (and also in Supplementary Figs. S2 and S3 in Supplementary Material for data sets D1 and D3) is that many more loci seem to support gene tree \(G_{6c}\) than \(G_{6a}\) or \(G_{6b}\), whereas the model predicts equal proportions for those three gene trees. This is in contrast to the simulated data set, 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 data set D1 (auto), assuming alternative species trees for M, S, and Y. The MLEs and log likelihood values are shown in Table 10. The (MSY) 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 \(\tau = 0\). Migration is detected only in the direction of S \(\rightarrow\) M when the assumed tree is (MSY). Note that our model assumes migration between the two in-group

### Table 10. MLEs and log likelihood values under M2 assuming different species trees for data set D1 (auto) of Table 4

<table>
<thead>
<tr>
<th>Species tree</th>
<th>(\tau_{\text{MSY}})</th>
<th>(\tau)</th>
<th>(\theta_{\text{MSY}})</th>
<th>(\theta_1)</th>
<th>(\theta_2)</th>
<th>(\theta_3)</th>
<th>(M_{12})</th>
<th>(M_2)</th>
<th>(\ell)</th>
</tr>
</thead>
<tbody>
<tr>
<td>((MS)Y)</td>
<td>21.3 ± 0.1</td>
<td>13.6</td>
<td>0.2 ((\tau_{\text{MSY}}))</td>
<td>40.0 ± 0.3</td>
<td>10.6 ± 0.3 ((\theta_{\text{MSY}}))</td>
<td>5.2 ± 0.6</td>
<td>12.7 ± 0.2</td>
<td>NA</td>
<td>0.0 ((M_{12}))</td>
</tr>
<tr>
<td>((MV)S)</td>
<td>10.7 ± 0.1</td>
<td>10.7</td>
<td>1.0 ((\tau_{\text{MSY}}))</td>
<td>53.5 ± 0.3</td>
<td>3.1 ± 0.3 ((\theta_{\text{MSY}}))</td>
<td>5.7 ± 0.4</td>
<td>8.2 ± 0.1</td>
<td>0.0 ((M_{12}))</td>
<td>0.0 ((M_2))</td>
</tr>
<tr>
<td>((SM)Y)</td>
<td>11.4 ± 0.1</td>
<td>11.4</td>
<td>0.1 ((\tau_{\text{MSY}}))</td>
<td>52.8 ± 0.3</td>
<td>3.1 ± 0.3 ((\theta_{\text{MSY}}))</td>
<td>11.3 ± 0.1</td>
<td>4.2 ± 0.3</td>
<td>0.0 ((M_{12}))</td>
<td>0.0 ((M_2))</td>
</tr>
</tbody>
</table>

Note: Estimates of \(\tau\), \(\theta\), and \(M\) are multiplied by 1000. Estimates of \(\tau_1\) and \(\tau_2\) hit the upper bound set in the program for trees ((MY)S) and (SY)M.

S2 and S3 in Supplementary Material. At the MLEs under M2 (Table 9, with \(\tau_{\text{MSY}} = 0.020\) and \(\tau_{\text{MS}} = 0.013\) fixed), the expected gene tree probabilities for any MSY locus are \(P(G_{6a}) = 0.1324\), \(P(G_{6b}) = 0.7368\), and \(P(G_{6c}) = P(G_{6d}) = P(G_{6e}) = 0.0436\), with the gene tree-species tree mismatch probability \(P(G_{6a}) + P(G_{6b}) = 0.0872\). Most loci have gene tree \(G_{6c}\) (Fig. 4), because the migration rate is low, so that \(G_{6c}\) 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 \(G_{6a}\), and are likely to have been transferred across species (from \(D.\) simulans to \(D.\) melanogaster since \(M_{\text{MS}} \approx 0\)). The top 41 loci, with \(P(G_{6a}) > 95\%\), are listed in Supplementary Table S8 in Supplementary Material. More than half of those loci were also inferred to have \(P(G_{6c}) > 95\%\) in the analysis of data set D4 (exons complete) (Supplementary Table S8 in Supplementary Material), suggesting that this inference was not very sensitive to the different filtering procedures applied to compile the data sets.
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 because 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 article, 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 data sets, 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 nonsister 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 data sets 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 data sets (see also Wang and Hey 2010). Should there be more than 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 (rs) and up to (2s– 1) population size parameters (bs). 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 article 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 data sets. 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 < τ1) 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 E1 into two segments: E1a: 0 < t < T and E1b: T < t < τ1. 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 hybridization may be modeled in the same framework (Twyford and Ennos 2011). Recent introgression or contamination may be modeled 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, whereas 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 data set similar to our data set 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 \( \theta_S = 0.0053 \) and \( \theta_M = 0.01352 \), which are close to our estimates from D1 (auto). The ancestral population size \( \theta_{MS} \) estimated by Wang and Hey ranges from 0.007 to 0.010, whereas our estimates are larger, at \( \theta_{MS} = 0.011 \) and \( \theta_{MSY} = 0.040 \). The M-S divergence time parameter is \( \tau_{MS} = 0.017 \) by Wang and Hey and 0.0136 in our analysis. A strong negative correlation between \( \theta_{MS} \) and \( \theta_{MSY} \) is expected in such analyses (Yang 2002). Wang and Hey (2010) estimated the migration rate (in our notation) to be \( M_{MS} = N_{M0} \theta_{MS} = 0 \) from D. melanogaster to D. simulans and \( M_{SM} = N_{M0} \theta_{SM} = 4.846 \times 0.00552/4 = 0.0067 \) from simulans to melanogaster. Our estimates under M2 are \( M_{MS} = 0 \) as well and \( M_{SM} = 0.0183 \), 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 data sets 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 data sets 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 data sets. 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 data sets, we conclude that the estimates obtained from those studies are largely consistent.

Different from Wang and Hey (2010), we also compiled and analyzed a data set for the X chromosome (D3 chrX) as well as two exon data sets: D4 (exons complete) and D5 (exons split). The use of multiple data sets, 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 data sets we analyzed support a model of unidirectional gene flow, from D. simulans to D. melanogaster, at the rate of \( \sim 0.02 \) migrant individuals per generation. We included the two exon data sets 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.

SUPPLEMENTARY MATERIAL

Supplementary material, including data files and/or online-only appendices, can be found in the Dryad data repository at http://dx.doi.org/10.5061/dryad.h0h4s.

ACKNOWLEDGMENTS

We thank Thomas Buckley, Bastien Boussau, and an anonymous reviewer for many critical and constructive comments, which have led to improvement of our article. We thank Bastien Boussau for the suggestion of inferring the gene loci that may have been transferred across species due to gene flow.

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 \( G_{1ac}, G_{1bc} \) and \( G_{1ab} \)) are identical. There is thus no need to keep track of the sequence IDs, even though the likelihood averages over all 18 gene trees (Supplementary Table S1 in Supplementary Material). Thus we consider a Markov chain with 8 states: 111, 112, 122, 222, 11, 12, 12, 11, 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.

$$f(G_2,t_0,t_1) = \frac{2}{P_2} e^{-\frac{t_0}{P_2}} + \sum_{j \in S_2} \left[ 3 \frac{2}{P_2} P_{3,111}^j(t_1) P_{11,j}(t_1-t_1) + \frac{2}{P_2} P_{3,112}^j(t_1) P_{12,j}(t_1-t_1) \right],$$

$$f(G_3,t_0,t_1) = \frac{2}{P_3} e^{-\frac{t_0}{P_3}} + \sum_{j \in S_3} \left[ 3 \frac{2}{P_3} P_{4,111}^j(t_1) P_{11,j}(t_1-t_1) + \frac{2}{P_3} P_{4,222}^j(t_1) P_{22,j}(t_1-t_1) \right],$$

$$f(G_4,t_0,t_1) = \frac{6}{P_4} e^{-\frac{t_0}{P_4}} + \sum_{j \in S_4} \left[ 6 \frac{2}{P_4} P_{5,111}^j(t_1) P_{11,j}(t_1-t_1) + \frac{2}{P_4} P_{5,222}^j(t_1) P_{22,j}(t_1-t_1) \right].$$

$$f(G_5,t_0,t_1) = \frac{6}{P_5} e^{-\frac{t_0}{P_5}} + \sum_{j \in S_5} \left[ 6 \frac{2}{P_5} P_{6,111}^j(t_1) P_{11,j}(t_1-t_1) + \frac{2}{P_5} P_{6,222}^j(t_1) P_{22,j}(t_1-t_1) \right].$$

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 (Supplementary Table S1 in Supplementary Material). 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 $1_12_2$, or $1_22_1$. 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 $1_11_2$, $1_21_2$, and $1_22_2$ are written as 111, 112, and 122, respectively. There are 21 states in the chain: 111, 112, 122, 211, 212, 222, 1_12_1_2, 1_22_2_1, 1_21_2_2, 1_22_2_1, 1_22_2_1, 1_22_2_1, 1_22_2_1, 1_22_2_1, 1_22_2_1, 1_22_2_1, 1_22_2_1, 1_22_2_1, 1_22_2_1, 1_22_2_1, 1_22_2_1, 1_22_2_1. The states of two sequences have the subscripts to indicate the sequence IDs. For example, $1_12_2$ means that sequences $b$ and $c$ have coalesced and their ancestor is in population 1, whereas sequence $a$ is in population 2.

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

$$f(G_{1e},t_0,t_1) = \frac{2}{P_1} P_{5,111}(t_1) \left( \frac{2}{P_1} P_{4,111}(t_1) + \frac{2}{P_1} P_{4,222}(t_1) + \frac{2}{P_1} P_{5,111}(t_1) \right).$$

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

$$f(G_{1b},t_0,t_1) = \frac{2}{P_1} P_{5,111}(t_1) \left( \frac{2}{P_1} P_{4,111}(t_1) + \frac{2}{P_1} P_{4,222}(t_1) + \frac{2}{P_1} P_{5,111}(t_1) \right).$$

For gene tree $G_2$, we have $t_1 < t_0, t_0 < t_1$, and

$$f(G_2,t_0,t_1) = \frac{2}{P_2} e^{-\frac{t_0}{P_2}} + \sum_{j \in S_2} \left[ \frac{2}{P_2} P_{3,111}^j(t_1) P_{11,j}(t_1-t_1) + \frac{2}{P_2} P_{3,112}^j(t_1) P_{12,j}(t_1-t_1) \right].$$
\[
f(G_{k,t_0},t_1) = \frac{2}{\theta_0} e^{-\frac{\theta_0}{\theta_t} t_1} \\
\times \sum_{k,j} \left[ \frac{2}{\theta_t} P_{k,111}(t_1) P_{j,k,1}(t_1-t_1) + \frac{2}{\theta_t} P_{k,211}(t_1) P_{j,k,2}(t_1-t_1) \right] \\
+ \frac{2}{\theta_t} P_{k,122}(t_1) P_{j,k,3}(t_1-t_1) + \frac{2}{\theta_t} P_{k,222}(t_1) P_{j,k,4}(t_1-t_1) \right].
\]

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 \( G_6 \) (Supplementary Table S1 in Supplementary Material). For tree shapes \( G_3 \) and \( G_5, \) the only gene tree possible is \( G_3 = \{a,b,c\}, \) whereas for the tree shape \( G_6, \) the three gene trees \( G_{6c}, \{a,b,c\}, G_{6c}(\{b,c\},a) \) and \( G_{6d}(\{c,a\},b) \) have the same prior density. Thus there is no need to trace the sequence IDs. There are 4 states in the chain: 113, 123, 223, 131/23, with the rate matrix given as follows.

\[
\begin{array}{cccc}
113 & 123 & 223 & 13/23 \\
\hline
113 & -2(\theta_{t1} + \theta_0) & 2 \theta_{t1} & 0 & c_1 \\
123 & -2(\theta_{t1} + \theta_0) & -2(\theta_{t1} + \theta_0) & 2 \theta_{t1} & 0 \\
223 & 0 & 2 \theta_{t1} & -2(\theta_{t1} + \theta_0) & c_2 \\
13/23 & 0 & 0 & 0 & 0
\end{array}
\]

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

\[
f(G_{k,t_0},t_1) = \frac{2}{\theta_0} e^{-\frac{\theta_0}{\theta_t} t_1} \times \sum_{j} P_{k,j}(t_1).
\]

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

\[
f(G_{6k},t_0,t_1) = \frac{2}{\theta_0} e^{-\frac{\theta_0}{\theta_t} t_1} \times \sum_{j} P_{k,j}(t_1) e^{-\frac{\theta_0}{\theta_t} t_1},
\]

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 \( \theta_k \) for \( t < t_0 \) or \( \theta_4 \) for \( t > t_0. \) There is no need to distinguish among \( G_1, G_3, \) and \( G_4, \) or between \( G_2 \) 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 \( G_4. \) For each one, we sum over three gene trees. Thus with initial state \( s = 333, \) we have

\[
f(G_{k,t_0},t_1) = \frac{2}{\theta_0} e^{-\frac{\theta_0}{\theta_t} t_1} \times \sum_{j} P_{k,j}(t_1), 0 < t_1 + t_0 < t_0 - t_1,
\]

where \( k = c,a, \) and \( b. \)
Similarly, for initial state $s = 133$ or $233$, we consider two tree shapes $G_2$ and $G_6$.

\[
f(G_2, t_0, t_1) = \begin{cases} 
\frac{1}{2} \beta^{t_2} - \frac{2}{2} \beta^{t_1} - \frac{2}{2} \beta^{t_0}, & t_1 < t_0, \\
\frac{2}{2} \beta^{t_2} - \frac{2}{2} \beta^{t_1} - \frac{0}{2} \beta^{t_0}, & 0 < t_1, t_0 < \infty, \\
\frac{0}{2} \beta^{t_2} - \frac{2}{2} \beta^{t_1} - \frac{2}{2} \beta^{t_0}, & t_0 < t_1 < \infty,
\end{cases}
\]

for $k = 3$.

\[(A.11)\]

**REFERENCES**


Obbard D. J., MacIennan I., Kim K. W., Rambaut A., O'Grady P. M.,
Jiggins F. M. 2012. Estimating divergence dates and substitution
Palmieri N., Nothe V., Chen J., Schlatterer C. 2014. Genome assembly
and annotation of a Drosophila simulans strain from Madagascar.
Patterson N., Richter D. J., Geurts S., Lander E. S., Reich D. 2006. Genetic
evidence for complex speciation of humans and chimpanzees.
Nature 441:1103–1108.
Rannala B., Yang Z. 2003. Bayes estimation of species divergence times
and ancestral population sizes using DNA sequences from multiple
Russo C. A., Takezaki N., Nei M. 1995. Molecular phylogeny and
Saitou N. 1988. Property and efficiency of the maximum likelihood
Sel S. G., Liang K.-Y. 1987. Asymptotic properties of maximum
likelihood estimators and likelihood ratio tests under nonstandard
Strobeck K. 1987. Average number of nucleotide differences in a sample
from a single subpopulation: a test for population subdivision.
Takahata N. 1988. The coalescent in two partially isolated diffusion
Takahata N., Satta Y., Klein J. 1995. Divergence time and population
Twyford A. D., Ennos R. A. 2011. Next-generation hybridization and
Wang Y., Hey J. 2010. Estimating divergence parameters with small
samples from a large number of loci. Genetics 184:363–379.
Wen D., Yu Y., Hahn M. W., Nakahide L. 2016. Reticulate evolutionary
history and extensive introgression in mosquito species revealed by
differentiation under various models of population structure.
Wilkinson-Herbots H. M. 2008. The distribution of the coalescence time
and the number of pairwise nucleotide differences in the "isolation
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
Yamamichi M., Gojobori T., Innan H. 2012. An autosomal analysis
gives no genetic evidence for complex speciation of humans and
Yang Z. 1994. Statistical properties of the maximum likelihood method
of phylogenetic estimation and comparison with distance matrix
Yang Z. 2002. Bayesian species delimitation using
2754.
Zhou W. W., Wen Y., Fu J., Xu Y. B., Jin J. Q., Ding L., Min M. S., Che J.,
Zhang Y. F. 2012. Speciation in the Rana chensinensis species complex
and its relationship to the uplift of the Qinghai-Tibetan Plateau. Mol.
Zhu T., Yang Z. 2012. Maximum likelihood implementation of an
isolation-with-migration model with three species for testing