


mstree: A Multispecies Coalescent Approach for Estimating Ancestral Population Size and Divergence Time during Speciation with Gene Flow

Junfeng Liu ^{1,*}, Qiao Liu², and Qingzhu Yang^{1,2,*}

¹Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

²Department of Automation, Tsinghua University, Beijing, China

*Corresponding authors: E-mails: liujunfeng@big.ac.cn; yangqz@mail.tsinghua.edu.cn.

Accepted: April 27, 2020

Abstract

Gene flow between species may cause variations in branch length and topology of gene tree, which are beyond the expected variations from ancestral processes. These additional variations make it difficult to estimate parameters during speciation with gene flow, as the pattern of these additional variations differs with the relationship between isolation and migration. As far as we know, most methods rely on the assumption about the relationship between isolation and migration by a given model, such as the isolation-with-migration model, when estimating parameters during speciation with gene flow. In this article, we develop a multispecies coalescent approach which does not rely on any assumption about the relationship between isolation and migration when estimating parameters and is called *mstree*. *mstree* is available at <https://github.com/liujunfengtop/MStree/> and uses some mathematical inequalities among several factors, which include the species divergence time, the ancestral population size, and the number of gene trees, to estimate parameters during speciation with gene flow. Using simulations, we show that the estimated values of ancestral population sizes and species divergence times are close to the true values when analyzing the simulation data sets, which are generated based on the isolation-with-initial-migration model, secondary contact model, and isolation-with-migration model. Therefore, our method is able to estimate ancestral population sizes and speciation times in the presence of different modes of gene flow and may be helpful to test different theories of speciation.

Key words: coalescent, gene tree, mathematical inequalities.

Introduction

The role of gene flow in speciation is a fundamental issue in evolutionary biology. Allopatric speciation considers complete lack of gene flow as prerequisite to the formation of new species. However, parapatric and sympatric speciation allow gene flow during speciation. Although allopatric speciation has been historically taken as the paramount mode of speciation (Futuyma and Mayer 1980), theoretical modeling and empirical evidence increasingly support that speciation can occur with gene flow (Gourbiere and Mallet 2010; Smadja and Butlin 2011; Feder et al. 2012).

There are usually two kinds of models to make inferences about gene flow during speciation. Some methods are based on an isolation-with-migration (IM) model (Wang and Hey 2010; Tian and Kubatko 2016; Dalquen et al. 2017) and

others on an isolation-with-initial-migration (IIM) model (Mailund et al. 2012; Costa and Wilkinson-Herbots 2017). However, the above two models include an assumption about the relationship between isolation and migration. Here, we use the properties of coalescent-based model in gene tree data for estimating the important parameters such as ancestral population sizes and divergence times without any assumption of the relationship between isolation and migration. Furthermore, we conduct simulations to examine the accuracy of the estimates of parameters. The simulation results show that our method can accurately estimate the parameters. At last, we compared *mstree* with the program 3s (Dalquen et al. 2017) and IMA3 (Hey et al. 2018) with simulation data; the simulation results show that *mstree* is faster than 3s and IMA3.

Materials and Methods

The Theoretical Model

Consider two closely related species (1 and 2) with an out-group species 3. We assume that there is only gene flow between two closely related species (fig. 1). We use τ_0 and τ_1 to denote the two species divergence times, scaled by mutation rate. Let $\theta_0 = 4N_0\mu$ and $\theta_1 = 4N_1\mu$ measure the two ancestral species population sizes. Here, μ is the mutation rate per site and generation, and the N_0 and N_1 denote the effective population sizes. There are five possible gene trees for a locus with three sequences (k , l , and m), which are from species 1, species 2, and species 3, respectively (fig. 2). For any locus, t_0 is the coalescent time among three sequences and t_1 is the coalescent time between two sequences. In the presence of gene flow between species 1 and species 2, the gene tree G_1 may be possible at a locus. Otherwise, only gene trees

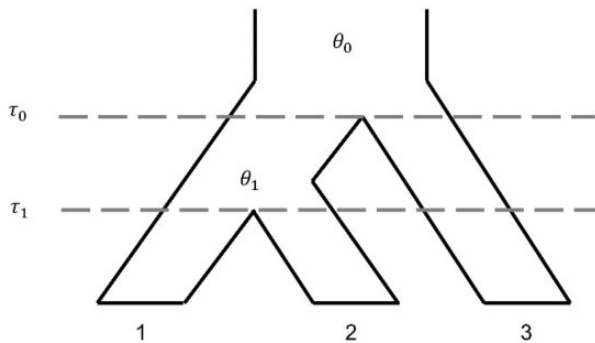


FIG. 1.—Species tree ((1, 2), 3) for three species. The species divergence times are denoted as τ_0 and τ_1 . The ancestral species population sizes are denoted as θ_0 and θ_1 .

G_2 – G_5 are possible. In this study, the term “loci” refers to independent or loosely linked short segments of the genome, and we assume that there is no recombination within a locus while different loci are free recombining. For tens of thousands of loci, there are some mathematical inequalities among the species divergence time, the ancestral species population size, and the number of gene trees, of which t_1 is larger than τ_1 . Based on the coalescent theory with no gene flow under given species, the probability of gene tree, of which t_1 belongs to $[\tau_1, \tau_0)$, is $1 - e^{-2(\tau_0 - \tau_1)/\theta_1}$; and the probability of gene tree, of which t_0 belongs to $[\tau_0, \tau'_0)$ and t_1 is less than τ_0 , is $1 - e^{-2(\tau'_0 - \tau_0)/\theta_0}$. We use $g_i([a, b], [c, d])$ as the number of gene trees with category G_i (fig. 2) and with t_0 is in $[a, b]$ and t_1 in $[c, d]$ for $i = 1, 2, \dots, 5$. Moreover, to simplify notation, let g_i denote the number of gene trees with category G_i (fig. 2) for $i = 1, \dots, 5$. Then, the formulas of cases A and C are as follows:

Case A: If $\tau_1 \leq \tau'_1 < \tau_0$, then $\frac{g_3 + g_4 + g_5}{g_2([\tau_0, \infty], [\tau'_1, \tau_0]) + g_3 + g_4 + g_5} \approx e^{-2(\tau_0 - \tau'_1)/\theta_1}$ for the category G_2 – G_5 (fig. 2).

Case B: If $\tau_0 \leq \tau'_0 < \tau''_0$, then $\frac{g_1([\tau'_0, \infty], [0, \tau_0]) + g_2([\tau'_0, \infty], [0, \tau_0])}{g_1([\tau'_0, \infty], [0, \tau_0]) + g_2([\tau'_0, \infty], [0, \tau_0])} \approx e^{-2(\tau''_0 - \tau'_0)/\theta_0}$ for the category G_1 – G_2 (fig. 2).

Case C: $\frac{g_4 + g_5}{g_2 + g_3 + g_4 + g_5} \approx \frac{2}{3} e^{-2(\tau_0 - \tau_1)/\theta_1}$ for the category G_2 – G_5 (fig. 2).

The above gene tree distributions allow us to compute the parameters θ_0 , τ_0 , θ_1 , and τ_1 . The approach of estimating the parameters is called mstree and the strategies are as follows. First, we can estimate the value of τ_0 based on the fact that the shape of gene tree may be $((k, l), m)$, $((k, m), l)$ or $((l, m), k)$

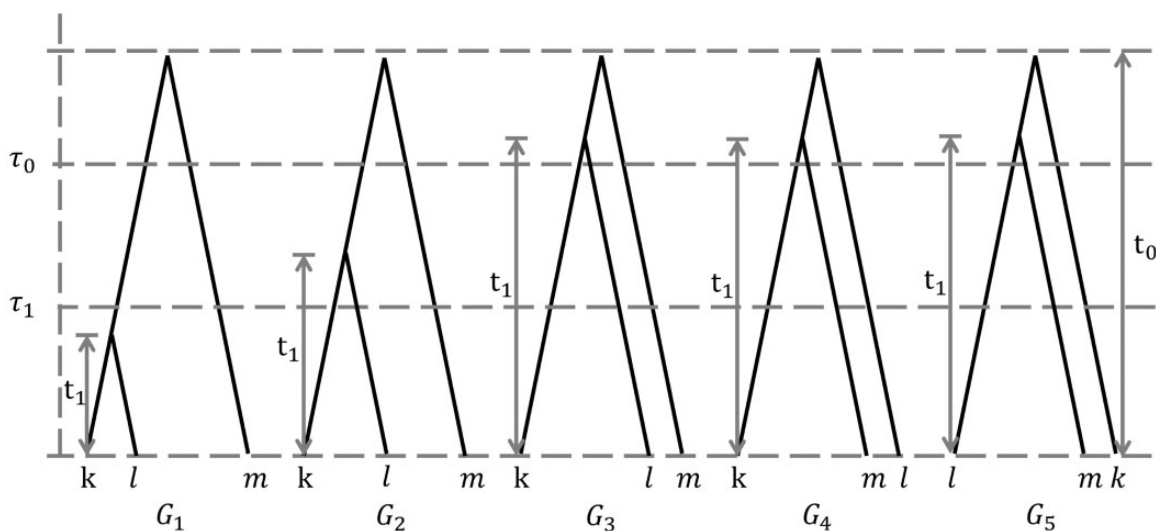


FIG. 2.—For three species (1–3) with gene flow between species 1 and 2, there are five categories of gene trees for any locus with three sequences (k , l , and m), which are from species 1, species 2, and species 3, respectively.

Table 1

The Estimated Species Divergence Time and Population Size with Different Threshold Value

| | Threshold | Hominoid | | | | Mangrove | | | |
|-----------|-----------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| | | θ_0 | θ_1 | τ_0 | τ_1 | θ_0 | θ_1 | τ_0 | τ_1 |
| IIM model | $\varepsilon = 0.007$ | 0.50 ± 0.02 | 0.49 ± 0.04 | 0.60 ± 0.00 | 0.42 ± 0.04 | 1.00 ± 0.06 | 0.97 ± 0.11 | 2.00 ± 0.01 | 1.25 ± 0.29 |
| | $\varepsilon = 0.01$ | 0.50 ± 0.02 | 0.49 ± 0.03 | 0.60 ± 0.00 | 0.41 ± 0.03 | 1.00 ± 0.05 | 0.98 ± 0.07 | 2.00 ± 0.01 | 1.14 ± 0.22 |
| | $\varepsilon = 0.03$ | 0.50 ± 0.02 | 0.50 ± 0.01 | 0.60 ± 0.00 | 0.39 ± 0.01 | 1.00 ± 0.02 | 0.99 ± 0.02 | 2.00 ± 0.01 | 0.99 ± 0.05 |
| SC model | $\varepsilon = 0.007$ | 0.50 ± 0.02 | 0.48 ± 0.05 | 0.60 ± 0.00 | 0.43 ± 0.05 | 1.00 ± 0.06 | 0.95 ± 0.14 | 2.00 ± 0.01 | 1.31 ± 0.33 |
| | $\varepsilon = 0.01$ | 0.50 ± 0.01 | 0.49 ± 0.04 | 0.60 ± 0.00 | 0.42 ± 0.03 | 1.00 ± 0.04 | 0.97 ± 0.09 | 2.00 ± 0.01 | 1.21 ± 0.27 |
| | $\varepsilon = 0.03$ | 0.50 ± 0.01 | 0.50 ± 0.02 | 0.60 ± 0.00 | 0.39 ± 0.01 | 1.00 ± 0.02 | 0.99 ± 0.03 | 2.00 ± 0.01 | 1.01 ± 0.07 |
| IM model | $\varepsilon = 0.007$ | 0.50 ± 0.02 | 0.48 ± 0.06 | 0.60 ± 0.00 | 0.43 ± 0.05 | 1.00 ± 0.06 | 0.95 ± 0.15 | 2.00 ± 0.01 | 1.34 ± 0.34 |
| | $\varepsilon = 0.01$ | 0.50 ± 0.01 | 0.49 ± 0.04 | 0.60 ± 0.00 | 0.41 ± 0.04 | 1.00 ± 0.04 | 0.96 ± 0.11 | 2.00 ± 0.01 | 1.22 ± 0.28 |
| | $\varepsilon = 0.03$ | 0.50 ± 0.01 | 0.50 ± 0.02 | 0.60 ± 0.00 | 0.39 ± 0.01 | 1.00 ± 0.02 | 0.99 ± 0.03 | 2.00 ± 0.01 | 1.00 ± 0.07 |

NOTE.—The hominoid set is $\theta_0 = \theta_1 = 0.005$, $\tau_0 = 0.006$, and $\tau_1 = 0.004$. The mangrove set is $\theta_0 = \theta_1 = 0.01$, $\tau_0 = 0.02$, and $\tau_1 = 0.01$. θ and τ estimates are scaled by 10^2 . Gene flow is symmetrical and the migration rate is 1. ε is the threshold value in mstree. The number of loci is 10,000. The number of replicates is 1,000. IIM, isolation-with-initial-migration; SC, secondary contact; IM, isolation-with-migration. The best estimates are marked in bold.

with equal probability when t_1 is larger than τ_0 . If there exists t' that satisfies $2a \approx n_2 + n_3$, where a is the number of gene trees, of which t_1 is larger than t' and the shape is $((k, l), m)$ (fig. 2: G_3); n_2 is the number of gene trees, of which the shape is $((k, m), l)$ (fig. 2: G_4); and n_3 is the number of gene trees, of which the shape is $((l, m), k)$ (fig. 2: G_5), t' can be considered as the estimated value of τ_0 . Second, we can estimate the value of θ_0 based on the estimated value of τ_0 by using the formula $e^{-2(\tau_0'' - \tau_0')/\theta_0} \approx \frac{a}{a+b}$ in case B. When we choose t' and t'' that satisfy $\tau_0 \leq t' < t''$, the estimated value of θ_0 approximates $-2(t'' - t')/\log(a/a + b')$, where a is the number of gene trees, of which t_0 is larger than t'' and t_1 is less than τ_0 ; b' is the number of gene trees, of which t_0 belong to $[t', t'')$ and t_1 is less than τ_0 . Similarly, we can also estimate the value of θ_1 based on the estimated value of τ_0 by using the formula $e^{-2(\tau_0 - \tau_1')/\theta_1} \approx \frac{a}{a+b'}$ in case A. If we choose t' that is less than τ_0 and assume that t' is larger than τ_1 when t' is closed to τ_0 , the estimated value of θ_1 approximates $-2(\tau_0 - t')/\log(a/a + b')$, where a is the number of gene trees, of which t_1 is larger than τ_0 and b' is the number of gene trees, of which t_1 belongs to $[t', \tau_0)$. Lastly, we estimate the value of τ_1 based on the values of τ_0 and θ_1 by using the formula $\frac{2}{3}e^{-2(\tau_0 - \tau_1)/\theta_1} \approx \frac{n_2 + n_3}{a + n_2 + n_3}$ in case C. If there exists t' that is less than τ_0 and satisfies $\frac{2}{3}e^{-2(\tau_0 - t')/\theta_1} \approx \frac{n_2 + n_3}{a + n_2 + n_3}$, where a is the number of gene trees, of which t_1 is larger than t' and the shape is $((k, l), m)$ (fig. 2: G_2 – G_3); n_2 is the number of gene trees, of which the shape is $((k, m), l)$ (fig. 2: G_4); and n_3 is the number of gene trees, of which the shape is $((l, m), k)$ (fig. 2: G_5), t' can be considered as the estimated value of τ_1 .

The Simulation

We simulated gene trees by using the program ms (Hudson 2002) and converted gene trees to sequence data under JC69

model by using seq-gen (Rambaut and Grassly 1997). The example of command is as follows:

```
./ms id="465" 3 50000 -T -l 3 1 1 1 -m 1 2 0 -m 2 1 0 -em 0.667 1 2 4 -em 0.667 2 1 4 -em 1 1 2 0 -em 1 2 1 0 -ej 1 2 1 -ej 2 3 1 | tail -n + 4 | grep -v//> tree
```

```
./seq-gen -m HKY -l 500 -s 0.01 -t 2.0 < tree > infile
```

Compared with IMA3 and Analyzed Real Data

The model estimated by IMA3 is IM model, and we used a fixed true species topology for IMA3. The real data are the genomic sequences of the human (H), chimpanzee (C), and gorilla (G) from Burgess and Yang (2008). The data set comprises 14,663 autosomal loci, and the mean locus length is 508 bp.

Results

The Accuracy of mstree

We used the program ms (Hudson 2002) to simulate gene trees at multi loci under the IIM, secondary contact (SC), and IM model. For the IIM model, the gene flow stopped at $\frac{2}{3}\tau_1$ in the past; For the SC model, the time of SC began at $\frac{1}{3}\tau_1$ in the past. Two sets of parameter values were used, roughly based on estimates from the hominoids (Burgess and Yang 2008) and the mangroves (Zhou et al. 2007). They are as follows: $\theta_0 = \theta_1 = 0.005$, $\tau_0 = 0.006$, and $\tau_1 = 0.004$ (hominoids); $\theta_0 = \theta_1 = 0.01$, $\tau_0 = 0.02$, and $\tau_1 = 0.01$ (mangroves). For the three models, gene flow is symmetrical and the migration rate (the expected number of migrants per generation) is 1. The number of loci is 10,000 and the number of replicates is 1,000. Analyzing the simulation data by using mstree, the results show that the parameter estimates are very close to the true values and are not sensitive to the model's assumption about the relationship between isolation and migration (table 1). ε in table 1 is the threshold value in mstree and

Table 2

The Estimated Species Divergence Time and Population Size with Larger and Different Parameter Values

| | Threshold | $\theta_0 = 0.02, \theta_1 = 0.03, \tau_0 = 0.06, \tau_1 = 0.04$ | | | | $\theta_0 = 0.02, \theta_1 = 0.01, \tau_0 = 0.02, \tau_1 = 0.01$ | | | |
|-----------|----------------------|--|-----------------|-----------------|-----------------|--|-----------------|-----------------|-----------------|
| | | θ_0 | θ_1 | τ_0 | τ_1 | θ_0 | θ_1 | τ_0 | τ_1 |
| IIM model | $\varepsilon = 0.03$ | 2.00 ± 0.08 | 2.97 ± 0.10 | 5.99 ± 0.03 | 3.90 ± 0.17 | 2.00 ± 0.06 | 1.00 ± 0.02 | 2.00 ± 0.01 | 1.00 ± 0.03 |
| SC model | $\varepsilon = 0.03$ | 2.01 ± 0.06 | 2.95 ± 0.17 | 5.99 ± 0.03 | 4.03 ± 0.19 | 2.00 ± 0.06 | 1.00 ± 0.02 | 2.00 ± 0.02 | 1.00 ± 0.04 |
| IM model | $\varepsilon = 0.03$ | 2.00 ± 0.05 | 2.94 ± 0.29 | 5.98 ± 0.05 | 3.89 ± 0.46 | 2.00 ± 0.05 | 1.00 ± 0.02 | 2.00 ± 0.02 | 1.00 ± 0.04 |

NOTE.— θ and τ estimates are scaled by 10^2 . Gene flow is symmetrical and the migration rate is 1. ε is the threshold value in mstree. The number of loci is 10,000. The number of replicates is 1,000. IIM, isolation-with-initial-migration; SC, secondary contact; IM, isolation-with-migration.

describes the degree of approximation between two sides of the formulas in cases A, B, and C. For example, in case A, $\varepsilon = 0.03$ means the value of $e^{-2(\tau_0 - \tau_1')/\theta_1} - \frac{a}{a+b'}/\left(\frac{a}{a+b'}\right)$ should be <0.03 when $e^{-2(\tau_0 - \tau_1')/\theta_1} \approx \frac{a}{a+b'}$. In mstree, the value of ε must be <0.05 . The results in [table 1](#) show that the smaller ε increased the standard deviations of the parameter estimates and the estimate of τ_1 . Therefore, we suggest that the value of ε should be 0.03 when using mstree. Furthermore, we applied mstree to additional two parameter sets ([Dalquen et al. 2017](#)), which have larger parameter values and different values for two θ s ([table 2](#)). The results show that mstree still performs well.

The Factors That Influence Parameter Estimates

In addition, we performed more simulations to test how different factors influence parameter estimates, such as the number of loci, migration rate, and the direction of migration. The numbers of loci are 5,000, 10,000, and 50,000; the migration rates are 0.1, 1, and 10; and the directions of migration are symmetrical and asymmetrical. The results are shown in [supplementary tables S1–S6, Supplementary Material](#) online. For the parameters θ_0 , θ_1 , and τ_0 , the results show that larger number of loci makes the estimates more accurate and the estimates are not sensitive to the model's assumption, migration rate, and the direction of migration. For the parameter τ_1 , we have the same conclusion except for the case that migration rate is 10. When migration rate is 10, asymmetrical gene flow decreases the accuracy of τ_1 estimate. This indicates that the estimate of τ_1 is sensitive to the direction of migration with large migration rate ([supplementary fig. S1, Supplementary Material](#) online). The examples of above simulation commands are in the [supplementary file S2, Supplementary Material](#) online.

Compared with 3s and IMA3

The input file of mstree is gene tree, which is in Newick format, and the gene tree can be estimated from the observed sequence alignments where there must be three sequences, with one sequence from each species, at each locus.

Therefore, we need a program, such as PHYLIP, to infer gene trees when applying mstree to experimental data. Because inference of gene trees is associated with error and uncertainty, we did some simulations to investigate the effect of the gene tree uncertainty ([supplementary table S7, Supplementary Material](#) online). We used program ms and seq-gen ([Rambaut and Grassly 1997](#)) to generate sequence data under JC69 model and used program dnamlk in PHYLIP package to infer gene trees. Although there has been some decline in the accuracy of parameter estimates because of the inferred error of gene trees, the estimates of mstree are still near to the true values and not sensitive to the model's assumption. Comparing mstree with the program 3s ([Dalquen et al. 2017](#)) and IMA3 ([Hey et al. 2018](#)), mstree is faster than 3s and IMA3 ([supplementary table S7, Supplementary Material](#) online). Although 3s and IMA3 performed very well on some parameter estimates, the τ_1 estimates of 3s and τ_0 estimates of IMA3 were very poor.

Robustness of mstree and Analyzing Real Data

Though our method is not affected by gene flow between the sister species, our method assumes that there is no gene flow between the ingroup and the outgroup. Therefore, we examined the robustness of our method in the presence of gene flow between the ingroup and the outgroup. The results are shown in [supplementary table S8, Supplementary Material](#) online. Our method is robust to the simulations based on IIM model between the ingroup and the outgroup. For the simulations based on SC and IM model between the ingroup and the outgroup, the accuracy of parameter estimates is on the decline. At last, we apply mstree to the genomic sequences of the human (H), chimpanzee (C), and gorilla (G) ([Burgess and Yang 2008](#)). The estimates of parameters are similar to those of [Burgess and Yang \(2008\)](#), but the estimate of τ_{HC} is slightly higher ([supplementary table S9, Supplementary Material](#) online). In order to quantify uncertainty in the estimates obtained, we resort to bootstrapping with 100 replicates. The averages and the standard errors of estimates are as follows: $\hat{\theta}_{HCG} = 0.0032 \pm 0.0000$, $\hat{\theta}_{HC} = 0.0068 \pm 0.0005$, $\hat{\tau}_{HCG} = 0.0059 \pm 0.0000$, and $\hat{\tau}_{HC} = 0.0038 \pm 0.0005$.

Discussion

Supplementary table S8, Supplementary Material online, shows the performance of mstree in the presence of gene flow with species 3. Under the influence of gene flow between the ingroup and the outgroup, τ_0 was underestimated and was closed to the time that gene flow stopped except for IIM model. When τ_0 was underestimated, the estimates of other parameters were far away from the true value. Burgess and Yang (2008) estimated divergence times under the assumption of no gene flow. However, Zhu and Yang (2012) applied the test based on SIM3s model to a human–chimpanzee–gorilla genomic data and the test results suggested gene flow around the time of speciation of human and chimpanzee. Compared with the estimated divergence times from Burgess and Yang (2008), the analysis from mstree suggested migrations between sister species. In addition, there are two significant differences between mstree and COALGF (Tian and Kubatko 2016), which describes the distribution of coalescent histories under the coalescent model with gene flow: 1) mstree uses the coalescent history distribution under coalescent model without gene flow to infer model parameters based on summary statistics; however, COALGF computes probabilities of gene tree histories given species trees under the coalescent process with gene flow and the results obtained from COALGF may be used to infer model parameters based on a maximum likelihood framework. 2) mstree does not make any assumption about the mode of gene flow between sister taxa; however, COALGF assumes that the mode of gene flow between sister taxa is IM.

To summarize, we propose a multispecies coalescent approach, mstree, for estimating the parameters during speciation with gene flow. Theoretically, our method does not rely on any assumption about the relationship between isolation and migration. Furthermore, the simulation results demonstrate that mstree can accurately estimate species divergence time and ancestral population size regardless of IIM model, SC model, or IM model.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank David Bryant and three anonymous reviewers for many critical and constructive comments, which have led to improvement of our article. This work was supported by the National Natural Science Foundation of China (No. 31501081 to Q.Y.).

Literature Cited

- Burgess R, Yang Z. 2008. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Mol Biol Evol.* 25(9):1979–1994.
- Costa RJ, Wilkinson-Herbots H. 2017. Inference of gene flow in the process of speciation: an efficient maximum-likelihood method for the isolation-with-initial-migration model. *Genetics* 205(4): 1597–1618.
- Dalquen DA, Zhu TQ, Yang ZH. 2017. Maximum likelihood implementation of an isolation-with-migration model for three species. *Syst Biol.* 66(3):379–398.
- Feder JL, Egan SP, Nosil P. 2012. The genomics of speciation-with-gene-flow. *Trends Genet.* 28(7):342–350.
- Futuyma DJ, Mayer GC. 1980. Non-allopatric speciation in animals. *Syst Zool.* 29(3):254–271.
- Gourbiere S, Mallet J. 2010. Are species real? The shape of the species boundary with exponential failure, reinforcement, and the “missing snowball”. *Evolution* 64(1):1–24.
- Hey J, et al. 2018. Phylogeny estimation by integration over isolation with migration models. *Mol Biol Evol.* 35(11):2805–2818.
- Hudson RR. 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* 18(2):337–338.
- Mailund T, et al. 2012. A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. *PLoS Genet.* 8(12):e1003125.
- Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci.* 13(3):235–238.
- Smadja CM, Butlin RK. 2011. A framework for comparing processes of speciation in the presence of gene flow. *Mol Ecol.* 20(24):5123–5140.
- Tian Y, Kubatko LS. 2016. Distribution of coalescent histories under the coalescent model with gene flow. *Mol Phylogenet Evol.* 105:177–192.
- Wang Y, Hey J. 2010. Estimating divergence parameters with small samples from a large number of loci. *Genetics* 184(2):363–379.
- Zhou R, et al. 2007. Population genetics of speciation in nonmodel organisms: I. Ancestral polymorphism in mangroves. *Mol Biol Evol.* 24(12):2746–2754.
- Zhu T, Yang Z. 2012. Maximum Likelihood Implementation of an Isolation-with-Migration Model with Three Species for Testing Speciation with Gene Flow. *Mol Biol Evol.* 29(10):3131–3142.

Associate editor: David T.E. Bryant