Recent advances in Bayesian inference of isolation-with-migration models

Yujin Chung*

Department of Applied Statistics, Kyonggi University, Suwon 16227, Korea

Isolation-with-migration (IM) models have become popular for explaining population divergence in the presence of migrations. Bayesian methods are commonly used to estimate IM models, but they are limited to small data analysis or simple model inference. Recently three methods, IMa3, MIST, and AIM, resolved these limitations. Here, we describe the major problems addressed by these three software and compare differences among their inference methods, despite their use of the same standard likelihood function.
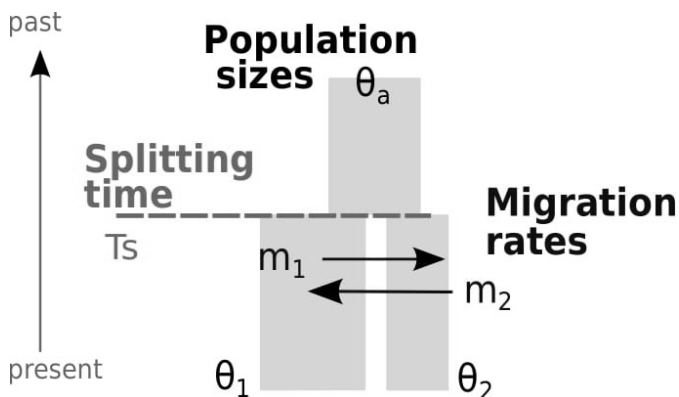
Keywords: bayesian analysis, coalescent theory, gene flow, isolation-with-migration model, phylogeny

Introduction

Divergence between populations and species has been a major interest in population genetics and evolution. Estimating divergence from genetic data is difficult because of conflicting evolutionary processes. Genetic drift elevates divergence between populations or between species, while gene flow can remove signals of divergence [1]. An isolation-with-migration (IM) model is a widely used demographic model describing the two conflicting signals. A typical 2-population IM model with six parameters (Fig. 1) depicts two populations (sizes θ_1 and θ_2, respectively) that arise from a single ancestral population (size θ_a) at time T_s in the past, while the two populations may exchange migrants at rates m_1 and m_2 ([1-3] for notations). Both population sizes and migration rates are assumed to be constant over time [4].

The challenges of inferring isolation models (with *no migrations*) and even phylogeny have been addressed by using a multispecies coalescent framework [5-11]. However, ignoring migrations can result in a biased estimation of splitting times of populations/species and may lead to a wrong phylogenetic tree estimation [12-16]. Efforts to distinguish between isolation and migration began about 20 years ago, and many methods have employed a Markov chain Monte Carlo (MCMC) simulation to infer an IM model [2-4,17-21]. However, most methods have a major roadblock of a long computational time of an MCMC simulation, which typically limits the amount of data that can be analyzed [1]. In addition, the joint estimation of both phylogeny and an IM model is known to be tremendously difficult [16].

Recently, three methods have been developed to address the scalability of the data and/or to jointly infer phylogeny in the presence of gene flow. IMa3 [16] is the most recent version of IM/IMa series software and infers the phylogeny and IM models. MIST [1] needs a known (or assumed) phylogeny but is able to analyze thousands of loci. AIM [14,22] is a package in the popular BEAST platform and also infers phylogeny in the presence of gene flow. Similar to other methods, these three methods implement the standard probabilistic framework and employ an MCMC simulation for inference.
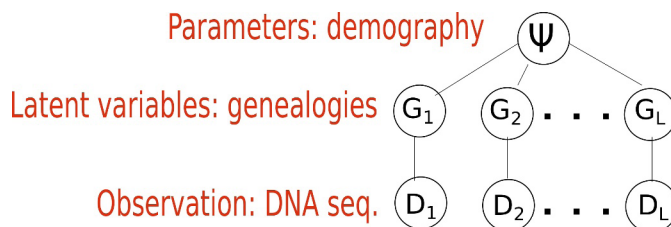
Fig. 1. Isolation-with-migration model with six demographic parameters.

Fig. 2. Standard model structure. Each locus has its own genealogy. Given genealogy, the genetic data and demography are assumed to be independent.

should not overlap or be closely located. Moreover, filtering using a four-gamete test [27] is essential to minimize potential recombination within a locus.

Standard Model Structure

When inferring an IM model from genetic data, the parameters of interest are demographic parameters of the IM model, denoted as a vector $\psi = (\theta_1, \theta_2, \theta_a, m_1, m_2, T_S)$. The ith locus D_i out of L loci are the observations, and the genealogy G_i of D_i is a latent variable that we cannot observe typically (Fig. 2). Fig. 2 depicts the structure of the standard models. The standard models address two levels of uncertainty: the distribution of DNA sequences given genealogy and that of genealogy given an IM model [11,12,25]. We typically assume that there is no recombination within a locus and free recombination between loci. In other words, the ith locus D_i out of L loci has as its own genealogy G_i and loci are independent. Given genealogy, the genetic data and demography $\psi = (\theta_1, \theta_2, \theta_a, m_1, m_2, T_S)$ are assumed to be conditionally independent. As the distribution of DNA sequences $p(D_i | G_i)$, diverse mutation or substitution models have been developed: infinite-site model [28], JC 69 model [29], HKY model [30], and GTR [31]. There are several useful methods for substitution model selection [25]. A coalescent process [32-34] is a well-known stochastic process for $p(G_i | \psi)$, the distribution of genealogy given a species tree or a demographic model. Most methods, including IMa3, MIST, and AIM, are based on this coalescent process. Based on this standard model structure, the likelihood function of ψ is built as follows:

$$\mathrm{L}(\psi) = \prod_{i=1}^{L} \int p(D_i | G_i) p(G_i | \psi) dG_i \qquad (1)$$

The likelihood function, so-called Felsenstein's equation [35], does not have a general closed-form and is difficult to numerically evaluate [3].

DNA Alignments

The advent of many inference methods is not commensurate to our skills of analysis using those programs. In order to ensure the use of appropriate programs and to correctly interpret results, it is essential to understand the inference methods used and the results that the programs provide. IMa3, MIST, and AIM all use similar standard probabilistic models and apply Bayesian inference, but their inference strategies and the types of results may be different. Therefore, users must first understand the differences in their inference methods.

To elucidate the current state of the art in the analysis of IM models, in this review article, we compare the three methods and accompanying software, IMa3, MIST, and AIM (BEAST platform). In particular, the data type and the underlying model structures will be discussed, followed by a brief summary of an MCMC algorithm and mixing issue. Then, this review article will focus on comparison of the advanced methods: IMa3, MIST, and AIM. We do not intend to explain the basic concepts of standard probabilistic models and MCMC algorithms, but extensive reviews of them are available elsewhere [9,23-25].

DNA Alignments

One of the most common types of data used in the analysis of IM models and phylogeny is DNA sequence alignments. Most methods, including IMa3, MIST, and AIM, assume the alignments are correct, although they are estimated from models of insertions and deletions [26]. The relatedness of homologous DNA sequences is considered to the result from past branching processes, so the DNA sequence alignments must be orthologs [22]. Moreover, no selection but a neutral evolution is assumed to act on alignments. Since most methods typically assume that there is no recombination within a locus and free recombination between loci, alignments

MCMC Simulation and the Mixing Problem

A feasible way to numerically evaluate the likelihood function (Eq. 1) is an MCMC simulation. Extensive reviews of fundamental concepts, diverse algorithms, and MCMC diagnosis are available elsewhere [23,25,36,37]. With a prior distribution on ψ, the posterior density of ψ given data is

$$p(\psi \mid D) \propto p(\psi) L(\psi) = p(\psi) \prod_{i=1}^{L} \int p(D_i \mid G_i) p(G_i \mid \psi) dG_i \qquad (2)$$

The target density of an MCMC simulation is

$P(\psi, G_1, \cdots, G_L \mid D) \propto p(\psi) \prod_{i=1}^{L} p(D_i \mid G_i) p(G_i \mid \psi)$, and a typical algorithm jointly simulates n samples from the target density:

$(\psi^1, G_1^1, \cdots, G_L^1), (\psi^2, G_1^2, \cdots, G_L^2), \cdots, (\psi^n, G_1^n, \cdots, G_L^n)$
$\sim P(\psi, G_1, \cdots, G_L \mid D)$

One of the benefits of such a simulation is an easy approximation of the marginal posterior density (Eq. 2) by making use of simulated values for the parameter of interest. For example, $\psi^1, \ldots \psi^n$, from the jointly simulated values

$(\psi^1, G_1^1, \cdots, G_L^1), (\psi^2, G_1^2, \cdots, G_L^2), \cdots, (\psi^n, G_1^n, \cdots, G_L^n),$

approximately follow $p(\psi \mid D)$ in Eq. (2) [38].

A popular MCMC algorithm is a Metropolis-Hastings within Gibbs sampling algorithm (Fig. 3). Within each iteration, all demographic parameters and genealogies are sequentially simulated. For example, Fig. 4A shows the state of the (t-1)th iteration for the genealogy of one locus and all demographic parameters ψ including splitting time T_s^{t-1}. If we try to update the splitting time at the tth iteration, we propose a new splitting time T_s^* using a proposal function q and either accept the new value $T_s^t = T_s^*$ with probability

$$\alpha = \min\left\{1, \frac{p(G^t \mid \psi^*) q(T_s^{t-1} \mid T_s^*)}{p(G^t \mid \psi^{t-1}) q(T_s^* \mid T_s^{t-1})}\right\}$$ or reject the new value and retain the previous state $T_s^t = T_s^{t-1}$ with 1-α, where ψ^* and ψ^{t-1} includes T_s^* and T_s^{t-1}, respectively.

1. Choose a starting value $G^0 = (G_1^0, \ldots, G_L^0)$ and Ψ^0.

2. At iteration t,

    i. For $i = 1, \ldots, L$, draw G_i^* from $q(G_i^* \mid G_i^{t-1})$. Take $G_i^t = G_i^*$ with probability α_1; $G_i^t = G_i^{t-1}$ with probability $1 - \alpha_1$, where q is a proposal function and

    $$\alpha_1 = \min\left\{1, \frac{p(D \mid G_i^*) p(G_i^* \mid \Psi^{t-1}) q(G_i^{t-1} \mid G_i^*)}{p(D \mid G_i^{t-1}) p(G_i^{t-1} \mid \Psi^{t-1}) q(G_i^* \mid G_i^{t-1})}\right\}$$ is an acceptance probability.

    ii. For the ith parameter $(i = 1, \ldots, 6)$, draw Ψ_i^* from $q(\Psi_i^* \mid \Psi_i^{t-1})$. Take $\Psi_i^t = \Psi_i^*$ with probability α_2; $\Psi_i^t = \Psi_i^{t-1}$ with probability $1 - \alpha_2$, where

    $$\alpha_2 = \min\left\{1, \frac{p(G^t \mid \Psi_i^*) q(\Psi_i^{t-1} \mid \Psi_i^*)}{p(G^t \mid \Psi_i^{t-1}) q(\Psi_i^* \mid \Psi_i^{t-1})}\right\}, \Psi_i^{t-1}$$ is the vector Ψ^{t-1} with the jth element replaced by Ψ_j^t $(j < i)$, and Ψ_i^* is the vector Ψ_i^{t-1} with the ith element replaced by Ψ_i^*.

3. Repeat Step 2 many times.

Fig. 3. A typical Metropolis-Hastings within Gibbs sampling algorithm.

While samples via a traditional Monte Carlo method are independent, MCMC samplers generate autocorrelated draws because the current value is either a different value or the same as the previous. Strong autocorrelations slow down traversing the posterior space and take longer to produce independent-like samples $\psi^1, \ldots, \psi^n \sim p(\psi|D)$ [23,25]. This phenomenon is called a poor mixing of a Markov chain. Mixing issues affect the efficiency, and hence the computing time of an MCMC simulation. In the inference of IM models, poor mixing is a major roadblock to the analysis of genomic data or the co-inference of phylogeny [1,16]. For example, the state of genealogy and demographic parameters are given as Fig. 4A. If a new splitting time proposed at the next iteration is not compatible with the state of genealogy (Fig. 4B), then $p(G^t|\psi) = 0$ and the acceptance probability is zero. Therefore, the newly proposed value is automatically rejected, and the previous state should be sampled until a compatible value is proposed. In other words, the acceptance rate of the splitting time is governed by the state of genealogies and can be very small if a lot of loci are considered.
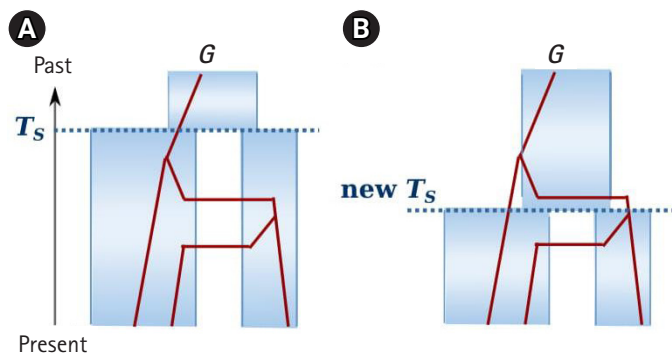
Fig. 4. An example of an Markov chain Monte Carlo step to update the splitting time T_S. (A) The current state of genealogy and all demographic parameters, including T_S. (B) A newly proposed splitting new T_S, which is not compatible with the state of genealogy.

Inference Methods

IMa3

The software series of IM/IMa were developed to infer IM models (Table 1) [39,40]. The first software, called IM, analyzes either a single locus [4] or multiple loci [2], and implements MCMC approaches to infer six demographic parameters $\psi = (\theta_1, \theta_2, \theta_a, m_1, m_2, T_S)$ of an IM model. In other words, the IM software simulates $(\psi, G_1, \ldots, G_L) \sim P(\psi, \ldots, G_L|D)$. Software IMa and IMa2 implement [3]. They simulate values of splitting time and genealogies

$$(T_S, G_1, \cdots, G_L) \sim p(T_S, G_1, \cdots, G_L|D)$$
$$\propto \prod_{i=1}^{L} p(D_i|G_i) p(G_1, \cdots, G_L|T_S) p(T_S),$$

but not population sizes and migration rate. It can be done by analytical integration of population sizes and migration rates:

$$\tag{3}$$
$$p(G_1, \cdots, G_L|T_S) = \int \cdots \int p(\theta_1) p(\theta_2) p(\theta_a) p(m_1) p(m_2)$$
$$\prod_{i=1}^{L} p(G_i|\psi) d\theta_1 d\theta_2 d\theta_a dm_1 dm_2$$

This yields a better mixing than software IM by reducing the number of parameters to sample, but it does not resolve the fundamental barrier of the relation between genealogies and splitting time. As a result of an MCMC simulation, the sampled values approximate the marginal posterior of the splitting time: $T_s^1, \ldots, T_s^n \sim p(T_s|D)$. Then IMa2 provides the posterior mean and the maximum *a posteriori* (MAP) estimate with highest posterior density intervals of the splitting time based on the marginal posterior density $p(T_s|D)$. To infer population sizes and migration rates, IMa and IMa2 do not simulate those parameter values, but directly approximate the marginal densities, $p(\theta_I|D)$ for i = 1, 2, a and $p(m_i|D)$ for i = 1, 2, from sampled values of $(T_S, G_1, \ldots, G_L)$. The approximated densities

Table 1. Comparison of Bayesian software MIST, AIM, and IMa3 (IM/IMa series)

Software	No. pop. to analyze	Inference method				Reference
		θ's and m's	Ts	Gᵃ	τ	
MIST	2	Density approx.[b]	Density approx.[b]	MCMC	No	[1]
AIM	2 or more	MCMC	MCMC	MCMC	MCMC	[14,22]
IMa3	2 or more	Density approx.[c]	MCMC	MCMC[d]	MCMC	[16]
IM	2	MCMC	MCMC	MCMC	No	[2,4]
IMa	2	Density approx.[c]	MCMC	MCMC	No	[3]
IMa2[e]	2 or more	Density approx.[c]	MCMC	MCMC	No	[3,17]

MIST, AIM, and IMa3 are compared in terms of the number of populations to analyze and inference methods by indicating what Markov chain Monte Carlo (MCMC) samples and which parameters' posterior densities are approximated rather than sampled. A similar comparison is made with IM/IMa series.
[a]All methods in this table sample genealogies and other mutation/substitution parameters from MCMC; [b]The joint posterior density of the 6 demographic parameters is approximated using MCMC samples; [c]The marginal posterior densities of the parameters are approximated using MCMC samples; [d]Hidden genealogies are sampled; [e]Variants: IMa2p [39] for parallel computation, IMGui [40] for any desktop OS.

$p(m_i|D)$ are employed to perform likelihood ratio tests (LRTs) from migration rates [3]. Software IMa infers 2-population IM models, but IMa2 extends IMa to infer multiple populations (see Table 1 for IMa2p and IMGui).

The most recent version called IMa3 modified the MCMC procedure of IMa2 to infer an IM model parameters as well as phylogeny [16], while IMa2 requires the phylogeny of multiple populations to be known. It is very difficult to co-estimate the phylogeny and IM model parameters, because sampling phylogeny together with IM parameters and genealogies also yields poor mixing. For example, a newly proposed phylogeny may not be compatible with the current state of migrations and is therefore rejected. IMa3 introduces pseudo-migrations, called *"hidden migrations,"* that occurred earlier than the splitting time so that a newly proposed splitting time or phylogeny is not instantly rejected but evaluated with non-zero acceptance probability. For example, if a newly proposed splitting time is younger than existing migrations (Fig. 4B), the migration paths older than splitting time are considered hidden migration paths (M_H) and the genealogy is the one without hidden migrations and compatible with the new splitting time. In other words, the current genealogy, given the new splitting time, is a so-called *"hidden genealogy"* $G_H = (G, M_H)$. Given phylogeny τ and demographic parameters ψ, the distribution of the hidden genealogy is partitioned into those of hidden migrations and the genealogy without hidden migrations: $p(G_H|\psi,\tau) = p(G|\psi,\tau)p(m_H|\psi,\tau)$. Therefore, in the presence of incompatible migration paths, a newly proposed splitting time or phylogeny is not automatically rejected. As a result, IMa3 simulates phylogeny, splitting times and hidden genealogies:

$$(\tau^1, T_S^1, G_{H,1}^1, \cdots, G_{H,L}^1), (\tau^2, T_S^2, G_{H,1}^2, \cdots, G_{H,L}^2), \cdots, (\tau^n, T_S^n, G_{H,1}^n, \cdots, G_{H,L}^n)$$
$$\sim P(\tau, T_S, G_{H,1}, \cdots, G_{H,L}|D) \propto p(\tau)p(\psi)\prod_{i=1}^{L} p(D_i|G_i)p(G_i|\psi,\tau)p(m_{H,i}|\psi,\tau)$$

Then $\tau^1, \dots, \tau^n$ from the MCMC samples approximately follow the marginal posterior $p(\tau|D)$. Similar to IMa2, demographic parameters are estimated based on their approximated marginal posteriors.

MIST

Software MIST [1] implements a 2-step analysis. First, it simulates genealogies without migrations (so-called *coalescent trees* λ) via an MCMC simulation. Note that no information about a demographic model is necessary in the first step, which alleviates the mixing problem. Second, the joint posterior density $p(\psi|D)$ in Eq. (2) is approximated from the sampled coalescent trees, and the MAP estimations of all demographic parameters are found.

Although MIST does not sample migrations and the underlying demographic model in step 1, the same posterior density $p(\psi|D)$

in Eq. (2) is inferred. It is done by separating migration paths from genealogies and applying the importance sampling [38]. The separation of migration paths enables the analytical computation of the density of a coalescent tree:

$$\mathrm{p}(\lambda_i|\psi) = \int \mathrm{p}(G_i|\psi)dM_i \qquad (4)$$

where the ith genealogy $G_i = (\lambda_i, M_i)$ and M_i is the set of all migration information. This rewrites Eq. (2) as follows:

$$\mathrm{p}(\psi|D) \propto p(\psi)\prod_{i=1}^{L} \int p(D_i|\lambda_i)p(\lambda_i|\psi)d\lambda_i \qquad (5)$$

The exact computation of $p(\lambda_i|\psi)$ employs a continuous time Markov chain representation [1]. In order to reduce the computational burden of the numerical integration in Eq. (5) $p(\lambda_i|\psi)$ by an MCMC simulation, the importance sampling method was employed. That is, MCMC samplers simulate coalescent trees from posterior $\check{p}(\lambda_i|D_i) \propto p(D_i|\lambda_i)\check{p}(\lambda_i)$ rather than $p(\lambda_i|\psi)$, where and $\check{p}(\lambda_i)$ is a flat prior. This MCMC simulation in step 1 does not use any information from the underlying IM model. The use of $p(\lambda i|D_i)$ rather than $p(\lambda_i|\psi)$ is compensated later in step 2 when the joint posterior density is approximated:

$$\mathrm{p}(\psi|D) \propto \mathrm{p}(\psi)\sum_{k=1}^{L} p(\lambda_k|\psi) \qquad (6)$$

As a result, MIST provides the MAP of all demographic parameters that maximize the joint posterior Eq. (6).

MIST has several strengths statistically and computationally. First, the computational complexity linearly increases with the number of loci. Analyses of thousands of loci do not give rise to mixing problems. Second, similar to IMa series, the approximate $p(\psi|D)$ in Eq. (6) can be used for LRTs for migration rates. While the IM/IMa series uses the marginal densities, MIST provides the joint distribution of all demographic parameters (Table 2). Since the estimations of demographic parameters are correlated, LRTs based on joint distributions have false-positive rates close to the expected value (e.g., 5%), even when very high false-positive rate occurred by LRTs based on marginal distributions [1,41,42]. Third, the importance sampling method enhances the computational efficiency for model comparisons. When different demographic models are compared, the simulated values from an MCMC simulation in step 1 can be repeatedly employed to infer different demographic models in step 2.

AIM

AIM [14] implements a Bayesian inference of phylogeny and IM models in using the BEAST platform [15,43]. BEAST is a software platform for phylogenetic analyses, phylodynamics, and population

Table 2. Prior assumptions, migration rate tests and scalability of Bayesian software MIST, IMa3, and AIM

		MIST	IMa3	AIM
Priors	θ's	Uniform	Uniform	Log-normal [44]
	m's	Uniform	Uniform	Exponetial[a]
	TS	Uniform	Uniform	Various [44]
	τ		Uniform	Various [44]
Tests for	m's	LRT[b]	LRT[c]	Bayes factor
Scalability[d]	Loci	Many (≤ 10K [1])	Moderate (≤ 200 [16])	Moderate (≤ 50 [14])
	Sequences	Few (≤ 8 [1])	Moderate (≤ 40 [16])	Moderate (≤ 133 [22])

[a]The prior for migration scalers; [b]Likelihood ratio test (LRT) compares the joint densities of the 6 demographic parameters; [c]LRT compares the marginal densities of migration rates; [d]The numbers of loci and sequences in the parentheses are collected from the cited studies. The scalability can depend on computer specifications as well.

genetics. starBEAST2 [44], an extended BEAST package, was added to estimate species trees in the absence of gene flow. AIM was recently added to estimate the posterior density $p(\psi \mid D)$ in Eq. (2) and $p(\tau \mid D)$, like IMa3. Similar to Chung and Hey [1], Müller et al. [22] drived a formula to compute the density of a coalescent tree $p(\lambda_i \mid \psi)$ a in Eq. (4) and additionally proposed approximations for a fast calculation. One approximation assumes the independence of lineages of the coalescent tree λ:

$$P_t(L_1 = l_1, L_2 = l_2 \mid \lambda, \psi) \approx P_t(L_1 = l_1 \mid \lambda, \psi) P_t(L_2 = l_2 \mid \lambda, \psi)$$

where L_1 and L_2 are lineages of λ at time t. AIM implements this independence approximation rather than the exact density $p(\lambda_i \mid \psi)$ in Eq. (4).

AIM reparamerized migration rates as follows: migration rate between populations A and B, $m_{A,B} = \alpha_{A,B} \frac{m_{tot}}{\delta_{AB}}$ where $\alpha_{A,B}$ is a scaler that is estimated between every pair of coexisting populations/species, δ_{AB} is the time to the most recent common ancestor from populations A and B coexisted, and m_{tot} is an estimated migration rate that allows for a prior distribution on the magnitude of the migration rate expected. This parameterization allows for smaller migration rates between more distant populations. Furthermore, each scaler $\alpha_{A,B} \sim Exp(1)$ and all scalers are assumed to be independent. AIM is able to use the priors previously implemented for species tree estimation in starBEAST2 [44].

AIM performs tests for migration rates based on Bayes factors (BFs) [14], while IMa3 and MIST use LRTs (Table 2). A BF as the ratio of marginal likelihoods [37] is wildely used for model selection. Since AIM is a package in the BEAST platform, users can take advantage of other existing packages and MCMC diagnostic tools. However, most packages in BEAST were developed independently [15,45]. Therefore, the results provided by different packages are not connected, and users need to be aware of the different terminologies by each package [15].

Discussion

IMa3, MIST, and AIM are advanced software that estimate demographic parameters of IM models. IMa3 and AIM sample population tree topologies and all or partial demographic parameters through an MCMC simulation. Therefore, their estimations are based on the marginal posterior distribution of parameters. MIST can estimate the joint posterior distribution of all parameters, thereby providing a joint estimation. IMa3 and AIM estimate population tree topology and migration rates, but their scalability to genomic data is limited or has not been yet examined. MIST scales well with genomic data and can be extended to infer population tree topologies. However, the software currently supports a joint estimation of demographic parameters of 2-population IM models.

While AIM uses BFs for migration rate test, IMa3 and MIST suggest LRTs. While IMa3 compares marginal posterior distributions, MIST provides joint posterior distributions for LRTs. When splitting times are recent, it is important to consider using joint distributions for LRTs in order to avoid a high false-positive [1,41,42].

Long-standing barriers to inferring IM models have been resolved by IMa3, MIST, and AIM. MIST can analyze genome-scale data without sever mixing problems in an MCMC simulation. IMa3 and AIM are able to estimate IM models and phylogeny in the presence of migrations. Nonetheless, there are still unresolved questions and no software implementing sophisticated models to answer the questions. One of the major interests for the future is to relax the strong assumption of constant migration rates and population sizes over time. Current methods that attempt to solve this problem are limited to small data or not capable of inferring IM models from real genetic data analysis [46,47].

ORCID

Yujin Chung: https://orcid.org/0000-0002-3801-3961

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

References

1. Chung Y, Hey J. Bayesian analysis of evolutionary divergence with genomic data under diverse demographic models. Mol Biol Evol 2017;34:1517-1528.

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

3. Hey J, Nielsen R. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc Natl Acad Sci U S A 2007;104:2785-2790.

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

5. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 2012;29:1969-1973.

6. Kubatko LS, Carstens BC, Knowles LL. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 2009;25:971-973.

7. Liu L. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 2008;24:2542-2543.

8. Liu L, Pearl DK. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst Biol 2007;56:504-514.

9. Liu L, Yu L, Kubatko L, Pearl DK, Edwards SV. Coalescent methods for estimating phylogenetic trees. Mol Phylogenet Evol 2009;53:320-328.

10. Rannala B, Yang Z. Efficient Bayesian species tree inference under the multispecies coalescent. Syst Biol 2017;66:823-842.

11. Degnan JH, Rosenberg NA. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol 2009;24:332-340.

12. Chung Y, Ane C. Comparing two Bayesian methods for gene tree/species tree reconstruction: simulations with incomplete lineage sorting and horizontal gene transfer. Syst Biol 2011; 60:261-275.

13. Leache AD, Harris RB, Rannala B, Yang Z. The influence of gene flow on species tree estimation: a simulation study. Syst Biol 2014;63:17-30.

14. Müller NF, Ogilvie HA, Zhang C, Drummond A, Stadler T. Inference of species histories in the presence of gene flow. Cold Spring Harbor: bioRxiv, Cold Spring Harbor Laboratory, 2018. Accessed 2019 Aug 3. Available from: https://doi.org/10.1101/348391.

15. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchene S, Fourment M, Gavryushkina A, et al. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput Biol 2019;15:e1006650.

16. Hey J, Chung Y, Sethuraman A, Lachance J, Tishkoff S, Sousa VC, et al. Phylogeny estimation by integration over isolation with migration models. Mol Biol Evol 2018;35:2805-2818.

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

18. Wakeley J, Hey J. Testing speciation models with DNA sequence data. In: Molecular Approaches to Ecology and Evolution (DeSalle R, Schierwater B, eds.). Basel: Birkhäuser, 1998. pp. 157-175.

19. Becquet C, Przeworski M. A new approach to estimate parameters of speciation models with application to apes. Genome Res 2007;17:1505-1519.

20. Dalquen DA, Zhu T, Yang Z. Maximum likelihood implementation of an isolation-with-migration model for three species. Syst Biol 2017;66:379-398.

21. Mailund T, Halager AE, Westergaard M, Dutheil JY, Munch K, Andersen LN, et al. A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. PLoS Genet 2012;8: e1003125.

22. Müller NF, Rasmussen DA, Stadler T. The structured coalescent and its approximations. Mol Biol Evol 2017;34:2970-2981.

23. Craiu RV, Rosenthal JS. Bayesian computation via Markov chain Monte Carlo. Annu Rev Stat Its Appl 2014;1:179-201.

24. Sousa V, Hey J. Understanding the origin of species with genome-scale data: modelling gene flow. Nat Rev Genet 2013;14:404-414.

25. Nascimento FF, Reis MD, Yang Z. A biologist's guide to Bayesian phylogenetic analysis. Nat Ecol Evol 2017;1:1446-1454.

26. Chatzou M, Magis C, Chang JM, Kemena C, Bussotti G, Erb I, et al. Multiple sequence alignment modeling: methods and applications. Brief Bioinform 2016;17:1009-1023.

27. Hudson RR, Kaplan NL. Statistical properties of the number of recombination events in the history of a sample of DNA sequenc-

es. Genetics 1985;111:147-64.

28. Kimura M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 1969;61:893-903.

29. Jukes TH, Cantor CR. Evolution of protein molecules. In: Mammalian Protein Metabolism (Munro HN, ed.). New York: Academic Press, 1969. pp. 21-132.

30. Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 1985;22:160-174.

31. Tavare S. Some probabilistic and statistical problems in the analysis of DNA sequences. Am Math Soc Lect Math Life Sci 1986;17:57-86.

32. Kingman JF. On the genealogy of large populations. J Appl Probab 1982;19:27-43.

33. Hudson RR. Properties of a neutral allele model with intragenic recombination. Theor Popul Biol 1983;23:183-201.

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

35. Felsenstein J. Phylogenies from molecular sequences: inference and reliability. Annu Rev Genet 1988;22:521-565.

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

37. Robert CP. Bayesian computational tools. Annu Rev Stat Its Appl 2014;1:153-177.

38. Robert CP, Casella G. Monte Carlo Statistical Methods. New York: Springer, 2004.

39. Sethuraman A, Hey J. IMa2p: parallel MCMC and inference of

ancient demography under the Isolation with migration (IM) model. Mol Ecol Resour 2016;16:206-215.

40. Knoblauch J, Sethuraman A, Hey J. IMGui-A Desktop GUI application for isolation with migration analyses. Mol Biol Evol 2017;34:500-504.

41. Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol Ecol 2014;23:3133-3157.

42. Hey J, Chung Y, Sethuraman A. On the occurrence of false positives in tests of migration under an isolation-with-migration model. Mol Ecol 2015;24:5078-5083.

43. Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, Xie D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol 2014;10:e1003537.

44. Ogilvie HA, Bouckaert RR, Drummond AJ. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Mol Biol Evol 2017;34:2101-2114.

45. Barido-Sottani J, Boskova V, Plessis LD, Kuhnert D, Magnus C, Mitov V, et al. Taming the BEAST-A community teaching material resource for BEAST 2. Syst Biol 2018;67:170-174.

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

47. Lan S, Palacios JA, Karcher M, Minin VN, Shahbaba B. An efficient Bayesian inference framework for coalescent-based non-parametric phylodynamics. Bioinformatics 2015;31:3282-3289.