Communications for Statistical Applications and Methods
2020, Vol. 27, No. 3, 385–395

https://doi.org/10.29220/CSAM.2020.27.3.385
Print ISSN 2287-7843 / Online ISSN 2383-4757

# A maximum likelihood approach to infer demographic models

Yujin Chung[1,a]

[a]Department of Applied Statistics, Kyonggi University, Korea

## Abstract

We present a new maximum likelihood approach to estimate demographic history using genomic data sampled from two populations. A demographic model such as an isolation-with-migration (IM) model explains the genetic divergence of two populations split away from their common ancestral population. The standard probability model for an IM model contains a latent variable called genealogy that represents gene-specific evolutionary paths and links the genetic data to the IM model. Under an IM model, a genealogy consists of two kinds of evolutionary paths of genetic data: vertical inheritance paths (coalescent events) through generations and horizontal paths (migration events) between populations. The computational complexity of the IM model inference is one of the major limitations to analyze genomic data. We propose a fast maximum likelihood approach to estimate IM models from genomic data. The first step analyzes genomic data and maximizes the likelihood of a coalescent tree that contains vertical paths of genealogy. The second step analyzes the estimated coalescent trees and finds the parameter values of an IM model, which maximizes the distribution of the coalescent trees after taking account of possible migration events. We evaluate the performance of the new method by analyses of simulated data and genomic data from two subspecies of common chimpanzees in Africa.

Keywords: demographic model, isolation-with-migration model, coalescent stochastic process, divergence time, evolutionary genomics

## 1. Introduction

In population genetics and molecular evolution, an isolation-with-migration (IM) model is a demographic model used to estimate the divergence time between two populations descended from their common ancestor population (Chung, 2019). For example, Figure 1(a) shows a 2-population IM model with six demographic parameters: $\Psi = (\theta_1, \theta_2, \theta_a, m_1, m_2, T_S)$, where the definition of each parameter is following (for the units of parameters, refer to Hey and Nielsen, 2004). The 2-population IM model describes a conceptual evolutionary history of two extant populations of sizes $\theta_1$ and $\theta_2$, respectively. Time $T_S$ ago, the two populations branched off from their common ancestor population of size $\theta_a$. The population sizes are assumed to be constant over time. After splitting, the two populations have independently evolved except for migrations between the two populations. It is assumed that migrations happened in both directions, but with different constant rates $m_1$ and $m_2$ (Figure 1(a)).

IM models are typically estimated from an alignment or aligned DNA sequences sampled from individuals of the extant populations of interest. An alignment is also called a *locus* because the aligned sequences are sampled from the same chromosome location. DNA sequences are aligned

[1] Department of Applied Statistics, Kyonggi University, Kwanggyosan-ro 154-42, Suwon, Gyeonggi-do 16227, South Korea. E-mail: yujinchung@kgu.ac.kr

Published 31 May 2020 / journal homepage: http://csam.or.kr

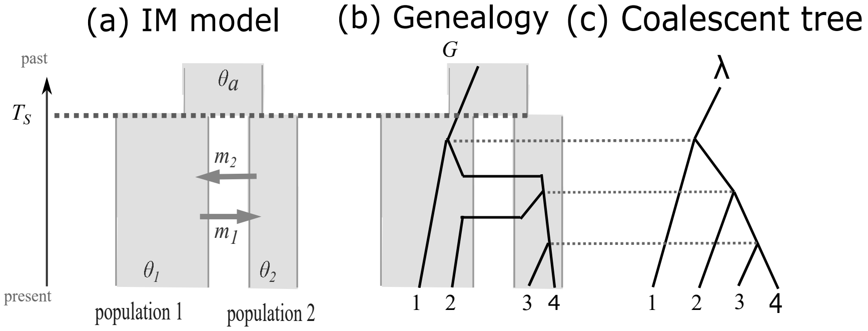

Figure 1: *(a) An example of an IM model. A 2-population IM model includes six demographic parameters: three population sizes ($\theta_1, \theta_2$ and $\theta_a$), two migration rates ($m_1$ and $m_2$), and a splitting time ($T_s$). (b) An example of a genealogy ($G$) given the IM model in (a). The genealogy depicts the evolutionary paths of four sequences sampled from the two extant populations. The genealogy includes three coalescent events (vertical paths of genes) and two migration events (horizontal paths). (c) The coalescent tree ($\lambda$) of the genealogy $G$ in (b). The coalescent tree includes the coalescent events on $G$, but not the migration events.*

so that they are descendants of a common ancestral sequence of the same length. Loci can have different evolutionary histories because of biological processes such as recombinations. Therefore it is common to assume that there is no recombination within loci and recombination unrestrictedly occurred between loci. In other words, loci are assumed to be statistically independent.

A *genealogy* ($G$) is a graphical representation of the evolutionary paths of an alignment from their common ancestral sequence (Figure 1(b)). A genealogy fundamentally includes vertical paths of how DNA sequences have descended generation by generation. For example, the genealogy in Figure 1(b) shows that lineages 3 and 4 branched off more recently than others. These branching patterns can also be explained backward in time. Two lineages coalesced into a single ancestral lineage backward in time, known as a *coalescent* event. Therefore, the genealogy in Figure 1(b) has three *coalescent* events. If the IM model in Figure 1(a) is the true demographic model, a genealogy of an alignment may include migration paths (horizontal paths) between the populations 1 and 2 on time period $(0, T_S)$. With the coalescent events of the genealogy in Figure 1(b), we can reconstruct a simpler tree in Figure 1(c), including no migration events. In this paper, the tree in Figure 1(c) is called the *coalescent tree* ($\lambda$) of genealogy in Figure 1(b). Since loci may have different histories, $G_i$ and $\lambda_i$ denote the genealogy and coalescent trees, respectively, of the $i^{th}$ locus of $n$ loci, for $i = 1, \dots, n$. Genealogies are latent variables because they cannot be observed. However, a genealogy plays an essential role in building a statistical model as a bridge connecting the genetic data to the demographic parameters.

To build the likelihood function of $\Psi$, the standard probability model takes account of the two levels of uncertainty (Felsenstein, 1988): (1) the probability distribution of an alignment given a genealogy using a mutation or substitution model (Kimura, 1969; Jukes and Cantor, 1969; Hasegawa *et al.*, 1985; Tavaré, 1986) and (2) the probability distribution of a genealogy given a demographic model using a stochastic process such as coalescent processes (Kingman, 1982; Hudson 1983). Then the probability of the $i^{th}$ locus given an IM model with parameters $\Psi$ can be obtained by integrating out all possible genealogies:

$$p(D_i|\Psi) = \int p(D_i|G_i)p(G_i|\Psi)dG_i, \tag{1.1}$$

where $p(D_i|G_i)$ is the probability of alignment $D_i$ given genealogy $G_i$ and $p(G_i|\Psi)$ is coalescent probability of genealogy $G_i$ given a demographic model with parameters $\Psi$. The likelihood function of all $n$ loci is as follows:

$$L_1(\Psi|\mathbf{D}) = \prod_{i=1}^{n} p(D_i|\Psi) = \prod_{i=1}^{n} \int p(D_i|G_i)p(G_i|\Psi)dG_i, \tag{1.2}$$

where $\mathbf{D} = (D_1, \ldots, D_n)$. The likelihood function is called Felsenstein's equation (Felsenstein, 1988).

The maximum likelihood estimation (MLE) of $\Psi$ is as follows:

$$\widehat{\Psi} = \arg\max_{\Psi} L_1(\Psi|\mathbf{D}). \tag{1.3}$$

The exact MLEs of demographic models are available for a limited case (Zhu and Yang, 2012). As far as we know, there is no general closed-form of the likelihood function (1.2), and it is extremely difficult to numerically compute likelihood values because of the vast space of a genealogy. Therefore, many Bayesian inferences have been developed by considering a prior distribution $p(\Psi)$ of $\Psi$. The posterior distribution of demographic parameters $\Psi$, given genetic data, is

$$p(\Psi|\mathbf{D}) \propto L_1(\Psi|\mathbf{D})p(\Psi).$$

Recently developed Bayesian methods mainly solved the mixing issues of an MCMC simulation, which was a long-standing barrier to analyses of genomic data from multiple populations (Hey *et al.*, 2018; Chung and Hey, 2017). Nonetheless, applying those methods to genomic analyses or demography inference for two or more populations is challenging because of the computational complexity (Chung, 2019).

We propose a new maximum likelihood approach, which is faster than the existing Bayesian methods while obtaining acceptable accuracy. The new method performs two steps. In step 1, the MLE of a coalescent tree for each locus is obtained. In step 2, the estimated coalescent trees are handled as data, and we use $p(\lambda|\Psi)$ as a likelihood function to find the MLE of $\Psi$. We evaluate the performance of the method in terms of accuracy, using simulated DNA sequences. To demonstrate the new method, we apply it to genomic data from two subspecies of common chimpanzees, *Pan troglodytes (P. t.) troglodytes* and *P. t. verus* (Prado-Martinez *et al.*, 2013).

## 2. A maximum likelihood approach

The new 2-step method is described for a 2-population IM model (Figure 1(a)) with six demographic parameters, $\Psi = (\theta_1, \theta_2, \theta_a, m_1, m_2, T_S)$. The Felsenstein's equation in (1.3) opened the door to the likelihood-based inference of population/species-level models from genetic data. The likelihood function incorporates two levels of uncertainties: (1) $P(D_i|G_i)$, a substitution probability and (2) $P(G_i|\Psi)$, a coalescent probability. However, it is difficult to evaluate the likelihood function, particularly the integration over genealogies. One reason is that the space of a genealogy is larger than that of a coalescent tree which grows double-exponentially with the number of sequences (Semple and Steel, 2003).

A genealogy is composed of the coalescent tree $\lambda$ and migration events $\mathcal{M}$, that is, $G = (\lambda, \mathcal{M})$, and genetic data is conditionally independent of migration events given a coalescent tree (Chung and Hey, 2017). In virtue of the property of genealogies, the substitution probability is $P(D_i|G_i) = P(D_i|\lambda_i)$, where $\lambda_i$ is the coalescent tree of $G_i$. Moreover, we can calculate the probability distribution of a coalescent tree given an IM model:

$$p(\lambda|\Psi) = \int p(G|\Psi)d\mathcal{M} = \int p(\lambda, \mathcal{M}|\Psi)d\mathcal{M}, \tag{2.1}$$

of which the integration over possible migration events is calculated using a continuous-time Markov chain representation of a genealogy (Zhu and Yang, 2012; Andersen *et al.*, 2014; Hobolth *et al.*, 2011; Chung and Hey, 2017) Then the locus-likelihood in (1.1) can be obtained by integrating over coalescent trees rather than genealogies:

$$L_1(\Psi|D_i) = p(D_i|\Psi) = \int p(D_i|\lambda_i)p(\lambda_i|\Psi)d\lambda_i. \tag{2.2}$$

The space of a coalescent tree is smaller than that of a genealogy, but remains vastly huge to compute the integration over possible trees.

We propose a two-step inference (Figure 2). In step 1, the new method uses the substitution probabilities to estimate the coalescent tree of each locus rather than integrating over coalescent trees:

$$\tilde{\lambda}_i = \arg\max_{\lambda_i} p(D_i|\lambda_i). \tag{2.3}$$

In step 2, the estimated trees, $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n$, are handled as data to build a likelihood function for $\Psi$:

$$L_2\left(\Psi|\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n\right) = \prod_{i=1}^{n} p\left(\tilde{\lambda}_i|\Psi\right). \tag{2.4}$$

We then propose the one that maximize the likelihood as the MLE of $\Psi$:

$$\widetilde{\Psi} = \arg\max_{\Psi} L_2\left(\Psi|\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n\right). \tag{2.5}$$

To execute the new method, we employ existing software each step. In step 1, estimated trees are obtained by using PAUP* (Swofford, 2002), popular software that provides maximum likelihood trees using diverse substitution models. In step 2, we use software MIST (Chung and Hey, 2017) which is a Bayesian approach to infer demographic models. Software MIST implements two-step Bayesian analyses using importance sampling. In step 1, for each locus, coalescent trees are sampled from a Markov chain Monte Carlo (MCMC) simulation. In step 2, the sampled trees are used to build the approximated posterior density function of $\Psi$, and the maximum a *posteriori* estimate (MAPE) of $\Psi$ is obtained. If the importance function used in step 1 of MIST is the posterior density of a coalescent tree corresponding to an improper flat prior, then the posterior density of $\Psi$ is approximated as follows:

$$p(\Psi|\mathbf{D}) \propto p(\Psi)\prod_{i=1}^{n}\left\{\sum_{j=1}^{k} p\left(\lambda_{i,j}|\Psi\right)\right\}, \tag{2.6}$$

where $\lambda_{i,j}$ is the $j^{th}$ sampled tree for locus $i$ from the MCMC simulation (Chung and Hey, 2017). If a single tree is sampled from an MCMC simulation ($k = 1$) and the prior distribution $p(\Psi)$ for independent parameters is constant over a wide range of $\Psi$, then the approximated posterior density is proportional to the likelihood function in (2.4):

$$p(\Psi|\mathbf{D}) \propto L_2\left(\Psi|\lambda_{1,1}, \ldots, \lambda_{n,1}\right). \tag{2.7}$$

If the input trees for step 2 of MIST are the estimated trees by PAUP*, then the MAPE of $\Psi$ under this circumstance (2.7) is the same as the MLE $\widetilde{\Psi}$ in equation (1.3). Therefore, to execute step 2 of the new method, we use the step 2 of MIST as if a single tree is sampled for each locus.

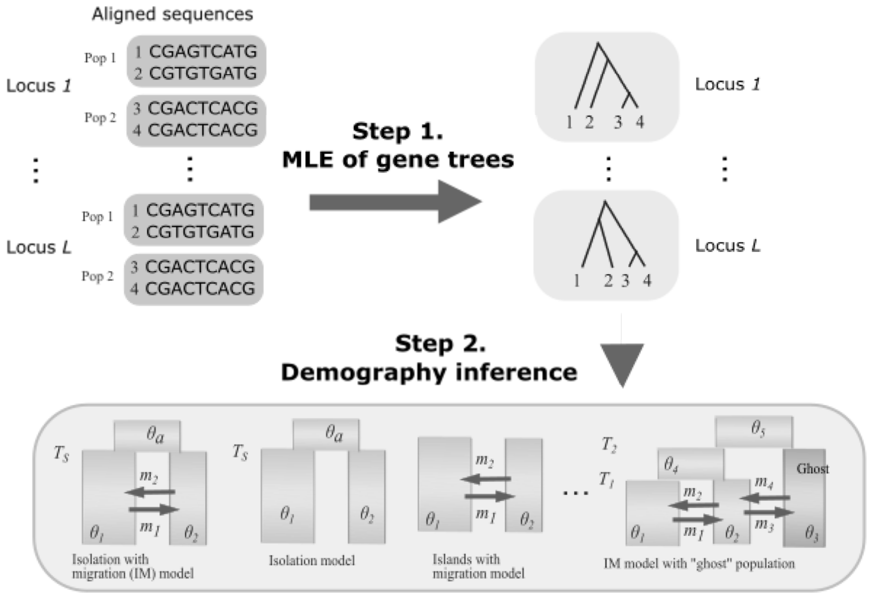

Figure 2: *The new method schematic. In step 1 the MLEs of coalescent trees for the aligned DNA sequences are obtained. In step 2, the set of estimated coalescent trees is used to infer a demographic model of interest. The same set of trees from step 1 can be used repeatedly to estimate different demographic models.*

The new maximum likelihood approach has several advantages. First of all, the new method scales well with the number loci and is expected to be faster than Bayesian approaches. The actual computing time of the optimization in step 2 depends on not only the number of trees to analyze but also the initial values, the curvature of the target function and etc. However, we can compare the computational complexity to calculate the target function for one iteration of an optimization algorithm. The step 2 of `MIST` maximizes the posterior function (2.6), and the time complexity to calculate the posterior function is $O(nk)$, where $n$ is the number of loci and $k$ is the number of the sampled trees for each locus. In other words, the computing time for each iteration is proportional to the total number of coalescent trees to analyze (Chung and Hey, 2017). The new method maximizes the likelihood function (2.4), where the computational complexity is $O(n)$. Therefore, the step 2 of the new method is expected to be faster than `MIST` as more coalescent trees are sampled from an MCMC simulation by `MIST`. Second, the estimated coalescent trees can be repeatedly used to estimate diverse demographic models (Figure 2). In step 1, coalescent trees are inferred from substitution models which are independent of a demographic model of interest. Once the trees are estimated, different demographic models can be inferred without repeating step 1. The third advantage is that the new method partially reduces a source of estimation variance by analytically integrating out possible migrations. However, the estimation errors of coalescent trees are added to the estimation errors of demographic parameters.

It is important to note that estimated genealogies are not appropriate to use. First, a genealogy with migration events cannot be defined without the given splitting times of populations. Second, migration events, particularly the times of migration events, are non-identifiable (Sousa *et al.*, 2011), and hence the credible intervals of migration times were very wide (Strasburg and Rieseberg, 2011). Therefore, most MCMC-based methods have to simulate genealogies and splitting times of populations together

Table 1: The mean squared errors of MLEs for each parameter

| Coalescent trees | No. loci | Parameter | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\theta_1$ | $\theta_2$ | $\theta_3$ | $m_1$ | $m_2$ | $T_S$ |
| True trees | 100 | 0.4145 | 0.0146 | 0.0615 | $1.94\times10^{-4}$ | $1.75\times10^{-3}$ | $6.79 \times 10^{-4}$ |
| | 1,000 | 0.0452 | 0.0011 | 0.0081 | $1.79\times10^{-5}$ | $1.15\times10^{-4}$ | $7.11\times10^{-6}$ |
| | 10,000 | 0.0048 | 0.0001 | 0.0006 | $1.34\times10^{-6}$ | $1.12\times10^{-5}$ | $1.14\times10^{-7}$ |
| ML trees | 100 | 0.5542 | 0.0166 | 19485.3 | 0.00038 | 0.00696 | 40.4721 |
| | 1,000 | 0.0675 | 0.0016 | 0.1606 | 0.00023 | 0.00066 | 0.0180 |
| | 10,000 | 0.0480 | 0.0003 | 0.1856 | 0.00017 | 0.00071 | 0.0171 |

The MLEs of the six parameters were obtained by analyzing (1) the true simulated coalescent trees and (2) the coalescent trees estimated from the simulated DNA alignments. MLE = maximum likelihood estimation.

(Hey and Nielsen, 2007; Hey *et al.*, 2018), which can lead to severe mixing problems (Chung, 2019).

## 3. Simulation

We evaluated the performance of the new maximum likelihood approach using simulations. First, we used a popular software in population genetics, called *ms* (Hudson, 2002). The software *ms* simulated coalescent trees from 2-populations IM model (Figure 1(a)) with $\theta_1 = 5$, $\theta_2 = 1$, $\theta_3 = 3$, $m_1 = 0.02$, $m_2 = 0.1$ and $T_S = 2$. The parameter values are the same as the simulation setting that Chung and Hey (2017) used to evaluate the performance of `MIST`. The number of loci varied as 100, 1,000, and 10,000, which is the same as the number of coalescent trees. Then we applied *seq-gen* (Rambaut and Grassly, 1997) to simulate DNA alignment of 1000 sites, given each of the simulated coalescent trees. We assumed a Jukes-Cantor substitution model (Jukes and Cantor, 1969) to simulate DNA sequences. For each case, 50 replicates were generated.

We first analyzed the simulated coalescent trees to evaluate the performance of the step 2 of the new method separately. Therefore, `MIST` was directly applied to the true simulated coalescent trees. Table 1 shows that the mean squared error (MSE) for each parameter is substantially reduced as more loci are analyzed. Both the bias and standard error of each demographic parameter decrease with the number of loci (data not shown).

The new approach analyzed the simulated DNA alignments. In step 1 of the analysis, `PAUP`* (Swofford, 2002) was applied to obtain the MLE of the coalescent trees for each locus. In step 2, `MIST` was applied to the estimated coalescent trees. Figure 3 shows the average of the MLEs for each parameter with standard errors. The standard error for each parameter was estimated by the standard deviation of the MLEs from 50 replicates. Figure 3 shows population sizes $\theta_1$ and $\theta_2$ from which DNA sequences were sampled tended to be underestimated, but the biases were small as −0.202 and −0.009, respectively, on 10,000-locus data. The ancestral population size $\theta_a$ was overestimated, and the bias was 0.428 on 10,000-locus data. The estimates for $m_1$ and $m_2$ were biased as 0.01243 and 0.0254, respectively on 10,000-locus data. The splitting time $T_S$ also tended to be underestimated with 1,000 or more loci (bias on 10,000-locus data: −0.126). Overall, the standard errors of all estimations were strictly decreasing with the number of loci. On the same set of demographic parameter values, the MAPEs by `MIST` were better than those by the new method (Figure 3 of Chung and Hey, 2017). The absolute bias and the standard errors of the MAPEs ranges 0.009–0.016 and 0.0028–0.0699, respectively, on 10,000-locus data (Chung and Hey, 2017). It is because the new method does not incorporate the uncertainty of coalescent tree estimations in Step 1, while `MIST` uses MCMC samples of coalescent trees. However, we note that the new method is faster than Bayesian approaches including `MIST`.

Some estimators have small biases; however, the overall accuracy in terms of MSEs for most

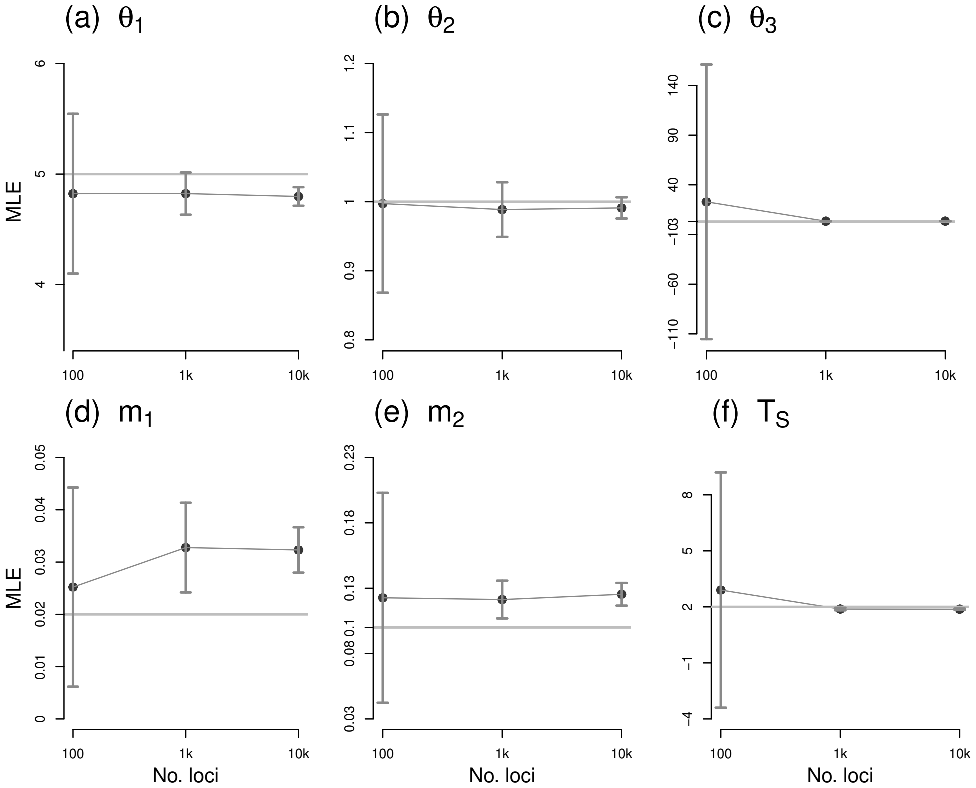

Figure 3: *The performance of the new method for a 2-population IM model. The average of the MLEs are plotted for each parameter whose true value is in gray horizontal line: (a)* $\theta_1 = 5$, *(b)* $\theta_2 = 1$, *(c)* $\theta_a = 3$, *(d)* $m_1 = 0.02$, *(e)* $m_2 = 0.1$, *and (f)* $T_S = 2$. *For each plot, vertical lines indicate standard errors, and the x axis for numbers of loci is on a log scale.*

parameters was better with the number of loci (Table 1). We note that the MSEs from the true tree analyses were induced by the step 2 of the method, and no errors were introduced from the step 1. Compared to the MSEs from the analysis of true trees, the MSEs from the analysis of DNA sequences were the errors accumulated through the two-step analysis of the new method. Therefore, the difference between two kinds of MSEs for each parameter is considered as the errors caused by using the point estimation of coalescent trees as input for step 2 and ignoring the estimation errors of the trees. Ignoring the uncertainty of tree estimation affects the estimations of $\theta_3$ and $T_S$ relatively more than other parameters. On 100-locus data, the MSEs for $\theta_3$ and $T_S$ from the analysis of DNA sequences were around $10^5$ and $10^4$ times larger than those from the true tree analysis, respectively. One possible reason is that the six parameters were jointly estimated and the estimate of $\theta_3$ is correlated with the estimate of $T_S$. Given a splitting time $T_S$, the information on coalescent trees such as the number of the coalescent events and coalescent times older than $T_S$ is sufficient statistics for $\theta_3$. Therefore, when a biased value of $T_S$ is chosen, a biased estimate of $\theta_3$ can be obtained as well. Migration rates, $m_1$ and $m_2$ are also highly correlated with $T_S$, but their MSEs were not as large as those for $\theta_3$ and $T_S$. It is possibly because the new method does not estimate individual migration paths, but uses the probability taking into account possible migration paths to estimate migration rates.

Table 2: Estimation of demographic parameters for two common chimpanzee subspecies

| | MLE | 90% CI | 95% CI |
|---|---|---|---|
| Population size of *P. t. troglodytes (Ptt)* | 1,056.73 | (755.13, 29,174.2) | (689.27, 30,264.6) |
| Population size of *P. t. verus (Ptv)* | 269.11 | (188.31, 350.47) | (187.11, 354.41) |
| Population size of common ancestor | 25,296.25 | (23,824.66, 63,146.33) | (19,139.05, 64,895.3) |
| Splitting time (years) | 8,103.96 | (6,589.36, 3,152,886) | (6,438.59, 3,204,850) |
| Migration rate per generation | | | |
| from *Ptt* to *Ptv* | 5.26e−4 | (1.29e−4, 7.36e−4) | (1.27e−4, 7.71e−4) |
| from *Ptv* to *Ptt* | 9.75e−4 | (6.44e−4, 0.002) | (6.19e−4, 0.002) |

The new method was applied to the genomic data of two chimpanzee subspecies, and the MLEs of the six demographic parameters were provided with 90% and 95% percentile confidence intervals (CIs) from bootstrappings. Estimates are shown on a demographic scale, using a per-site mutation rate per generation of 2e−8 and assuming 20 years per generation. Migration rates are backward in time.

## 4. Real data analysis

The new maximum likelihood approach was applied to the genomic data that Chung and Hey (2017) previously analyzed. The genomic data contains 1,000 loci of three sequences from each of two common chimpanzee subspecies, *Pan troglodytes troglodytes (Ptt)* from Central Africa and *Pan troglodytes verus (Ptv)* from West Africa. In step 1 of the new method, `PAUP`* was applied to each alignment to estimate the coalescent tree. Estimated trees with branches of length zero or with polytomies were disregarded, and 590 remaining trees were analyzed in the next step. In step 2, `MIST` was applied to 590 coalescent trees to infer 2-population IM models. To estimate confidence intervals (CIs) for each parameter, we performed a bootstrap method by resampling coalescent trees. The step 2 of the new method was applied to each of 100 bootstrapped data sets. Using the estimated parameter values from bootstrapped data, we calculated 90% and 95% percentile confidence intervals for each parameter.

The MLEs of demographic parameters by the new method were $\hat{\theta}_1 = 8.45\text{e}-5$ (*Ptt*), $\hat{\theta}_2 = 2.15\text{e}-5$ (*Ptv*), $\hat{\theta}_a = 0.002$, $\hat{T}_S = 8.1\text{e} - 6$, $\hat{m}_1 = 26289.9$, and $\hat{m}_2 = 48748.1$. Table 2 shows the estimated parameter values on an easy to interpret demographic scale. Converted parameters were effective population sizes, migration rates per generation, and the splitting times in years (see Hey and Nielsen, 2004; Chung and Hey, 2017 for the parameter conversion). To convert, we assumed a per-site mutation rate per generation of $2 \times 10^{-8}$ and assuming 20 years per generation (Chung and Hey, 2017). The 90% and 95% percentile CIs were wide except for the population size of *Ptv* (Table 2).

We compared our estimates to the previous studies of Bayesian analyses whose estimated values were similar (Won and Hey, 2005; Hey, 2010; Chung and Hey, 2017). Our estimates of sampled population sizes (*Ptt* and *Ptv*) and splitting time were smaller than those of the previous three studies. For example, the estimates by `MIST` were 27,081.38, 6,342.3, and 347,732, respectively (Chung and Hey, 2017). However, the bootstrap CIs except for *Ptv* contained the estimates by Bayesian analyses. The estimates of the common ancestral population size and two migration rates were later than previous Bayesian analysis, and their CIs did not include previous estimates. Overall, the pattern of bias by our new method was similar to the simulations.

The new estimates were different; however, their relative values were broadly consistent with the previous results by Chung and Hey (2017). First, the population size of *Ptt* is estimated to be four times larger than that of *Ptv*. Moreover, in population genetics, evolutionary independence can be assessed by the *population migration rate* ($2NM$), where $N$ is the effective population size, $M$ is the migration rate per sequence per generation, and both $N$ and $M$ are on a demographic scale (Wright, 1931). If $2NM < 1$, then the migration rate from another population is low relative to the rate of random genetic drift (neutral evolution). When $2NM > 1$, however, the migration rate is relatively high, and in population genetics, the divergence of the receiving population are more likely affected by the source population (Felsenstein, 1976). Our estimate of $2N_1M_1$ is 1.111 which is larger than the estimate, 0.525, of $2N_2M_2$. In the previous study by `MIST`, the estimate of $2N_1M_1$ was also larger than that of $2N_2M_2$, although both were smaller than our estimates ($2N_1M_1 = 0.878$ and $2N_2M_2 = 3.113\text{e}{-13}$) with one of them close to zero (Chung and Hey, 2017).

## 5. Conclusions

We presented a new maximum likelihood approach to estimate an IM model from genomic data. The new method is a 2-step analysis. In step 1, the MLEs of coalescent trees are obtained using substitution models. In step 2, the estimated trees were analyzed to find the MLEs of demographic parameters. The likelihood function for demographic parameters is the probability of coalescent trees after exactly accounting for all possible migration events. Removing migrations from inference enables us to estimate latent variables, coalescent trees. Compared to MCMC based methods that sample latent variables of genealogies or coalescent trees, the new method is computationally fast and scales with the number of loci.

From simulations and real data analysis, we evaluated the performance of the new method. The sampled population sizes and the splitting time tended to be underestimated, while migration rates and the ancestral population size were overestimated. However, the biases were small from the simulations, and the percentile CIs for most parameters included the previous results of Bayesian analyses. The relative relations of estimations were also consistent with previous results. As future work, it is important to study the asymptotic properties of the estimations $\widetilde{\Psi}$ in equation (2.5). Additionally, comparing $\widetilde{\Psi}$ with $\widehat{\Psi}$ in equation (1.3) would be interesting.

It is straightforward to extend the new approach to diverse demographic models for three or more populations. When three or more populations are considered, the population-level phylogenetic tree is also a parameter of interest. The computational complexity of the current MCMC based methods is a major roadblock to genomic analysis and to the inference of demographic models and the phylogeny of multiple populations (Chung, 2019). In particular, it is very challenging to estimate a demographic model along with the phylogenetic tree of multiple populations from genomic data (Hey *et al.*, 2018; Chung 2019). It is worth evaluating if the new method estimates a correct order of splitting times of three or more populations, although the estimations of splitting times could be biased. The topology of the population-level phylogenetic tree is easily reconstructed iIf we know the order of splitting times of multiple populations, then the topology of the population-level phylogenetic tree is easily reconstructed. Therefore, we expect that our method opens the door to a fast estimation of diverse IM models and the phylogenetic tree from genomes.

## Acknowledgements

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2018R1C1B5044541).

## References

Andersen LN, Mailund T, and Hobolth A (2014). Efficient computation in the IM model, *Journal of Mathematical Biology*, **68**, 1423–1451.

Chung Y (2019). Recent advances in Bayesian inference of isolation-with-migration models, *Genomics & Informatics*, **17**, e37.

Chung Y and Hey J (2017). Bayesian analysis of evolutionary divergence with genomic data under diverse demographic models, *Molecular Biology and Evolution*, **34**, 1517–1528.

Felsenstein J (1976). The theoretical population genetics of variable selection and migration, *Annual Review of Genetics*, **10**, 253–280.

Felsenstein J (1988). Phylogenies from molecular sequences: inference and reliability, *Annual Review of Genetics*, **22**, 521–565.

Hasegawa M, Kishino H, and Yano T (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA, *Journal of Molecular Evolution* **22**, 160–174.

Hey J (2010). The divergence of chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses, *Molecular Biology and Evolution*, **27**, 921–933.

Hey J, Chung Y, Sethuraman, A., Lachance J, Tishkoff S, Sousa VC, and Wang Y (2018). Phylogeny estimation by integration over isolation with migration models, *Molecular Biology and Evolution*, **35**, 2805–2818.

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

Hey J and Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics, *PNAS*, **104**, 2785–2790.

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

Hudson RR (1983). Properties of a neutral allele model with intragenic recombination, *Theoretical Population Biology*, **23**, 183–201.

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

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

Kimura M (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations, *Genetics*, **61**, 893–903.

Kingman JF (1982). On the genealogy of large populations, *Journal of Applied Probability*, **19**, 27–43.

Prado-Martinez J, Sudmant PH, Kidd JM, *et al.* (2013). Great ape genetic diversity and population history, *Nature*, **499**, 471–475.

Rambaut A and Grassly NC (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees, *Bioinformatics* **13**, 235–238.

Semple C and Steel M (2003). *Phylogenetics*, Oxford University Press, New York.

Sousa VC, Grelaud A, and Hey J (2011). On the nonidentifiability of migration time estimates in isolation with migration models, *Molecular Ecology*, **20**, 3956–3962.

Strasburg JL and Rieseberg LH (2011). Interpreting the estimated timing of migration events between hybridizing species, *Molecular Ecology*, **20**, 2353–2366.

Swofford D (2002). *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0*, Sinauer Associates.

Tavaré S (1986). Some probabilistic and statistical problems in the analysis of DNA sequences, *Lectures on Mathematics in the Life Sciences*, **17**, 57–86.

Won YJ and Hey J (2005). Divergence population genetics of chimpanzees, *Molecular Biology and Evolution*, **22**, 297–307.

Wright S (1931). Evolution in mendelian populations, *Genetics*, **16**, 97–159.

Zhu T and Yang Z (2012). Maximum likelihood implementation of an isolation-with-migration model with three species for testing speciation with gene flow, *Molecular Biology and Evolution*, **29**, 3131–3142.

*Received March 20, 2020; Revised April 14, 2020; Accepted April 20, 2020*