Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Evolutionary Bioinformatics

Approximate Likelihood Estimation of Divergence Time Range Using a Coalescent-based Model

Arindam RoyChoudhury
Department of Biostatistics, Columbia University.

ABSTRACT: We present an estimation of divergence time-range based on a coalescent model. This model sum the probability of coalescent trees, taking into account the effect of incomplete lineage sorting. Maximum likelihood estimate based on this model has been computed previously; however, a formula for divergence time-range estimate or confidence interval has never been presented for this model as the expression of the likelihood makes this estimation difficult to compute. Our formula for the divergence time-range estimate can be readily coded into a program. We did not use a simulation or resampling-based approach and therefore our method is fast and less computationally intensive. We demonstrate that our method is much faster and as accurate a simulation-based approach.

KEYWORDS: divergence time, range, confidence interval, maximum likelihood, coalescent

CITATION: RoyChoudhury. Approximate Likelihood Estimation of Divergence Time Range Using a Coalescent-based Model. *Evolutionary Bioinformatics* 2013:9 499–509 doi: 10.4137/EBO.S13080.

RECEIVED: August 25, 2013. **RESUBMITTED:** October 20, 2013. **ACCEPTED FOR PUBLICATION:** October 29, 2013.

ACADEMIC EDITOR: Jike Cui, Associate Editor

TYPE: Original Research

FUNDING: Author discloses no funding sources.

COMPETING INTERESTS: Author discloses no potential conflicts of interest.

CORRESPONDENCE: ar2946@columbia.edu

Introduction

Estimation of divergence time is an integral part of population genetics and evolutionary biology. It is a common practice[1,2,3] to estimate a range or a confidence interval of divergence time rather than a point-estimation (ie, a single value).

Despite being a common practice, maximum likelihood estimation of the time-range poses a challenge for coalescent-based models. This is partly due to the complicated expression of likelihoods of the coalescent-based models, and partly due to the fact that a likelihood-based formula for confidence interval requires the computation of estimated Fisher's information matrix, which has an even more complicated expression than the likelihood.

In this article, we present a divergence time range or a confidence interval based on the maximum likelihood estimation (MLE) of the divergence time from a coalescent-based model. We focus on the coalescent model of population evolution by Nielsen et al,[4] where the likelihood is computed by summing up over all possible coalescent trees between the present day and most recent common ancestor (MRCA). This one-step procedure takes into account the uncertainties of estimating the coalescent trees, as well as the effect of incomplete lineage sorting.

The coalescent framework is consistent with several models of population evolution, including a diffusion model, a Wright-Fisher model, a continuous-time, or a discrete-time Moran model (see[5,6]). Specifically, the coalescent framework in[4] requires a model with finite population size mating randomly within each population. The framework works with both monoecious and dioecious organisms and for both haploids and diploids. A complete separation between the populations is assumed at the point of divergence, and consequently, between-population mating are not allowed after the divergence. This also means that the coalescent events cannot take place between two individuals of different populations (between the present and the point of divergence).

It is further assumed that each locus has two alleles, and the allele-frequency spectrum for these biallelic loci is modeled with a symmetric Beta distribution as in.[5–9] This assumption of a symmetric Beta distribution results in a symmetric Beta-Binomial distribution for the allele-count. (Note that, the assumption of Beta distribution comes from a diffusion approximation; see, for example).[10]

The model is valuable for estimating the divergence time from related populations, as it produces the likelihood directly from the data, bypassing the gene trees. This is achieved by computing the exact probability of each coalescent tree as well as probabilities of allele frequencies given these trees and then summing over them with a closed-form mathematical formula. Thus, potential misestimation due to gene tree incongruence (eg, incomplete lineage sorting, see for example),[11] is avoided.

While computing the probabilities of the allele counts, it is assumed in this model that the effect of new mutations in allele frequencies between the MRCA and the present are negligible. This is an appropriate assumption for divergence estimation from related populations. This is because the divergence times in such populations are typically short, making the number of mutations very small. The variation in allele type is assumed to come from the mutations at or before the MRCA. A detailed mathematical description of the model probabilities is provided below.

There have been significant developments made in this model since its inception by[4] and a number of methods of inference have been proposed. Later[12] introduced a Markov Chain Monte Carlo-based on this model. More recently,[6] introduced a pruning algorithm to systematically compute the likelihood under this model for inference on a population tree. This approach builds on the approach of[4] and makes it possible to compute the likelihood of a large tree under this model. An innovative two-stage pruning algorithm is introduced for simultaneously keeping track of probability of the number of lineages, and allele counts among them. Later[7] introduced a composite likelihood method based on this model that can be used to analyze dependent data. This method treats the dependent allele counts from nearby loci as independently by multiplying the marginal likelihoods obtained from each locus. As a result, a composite likelihood is computed, which is then maximized to obtain a maximum composite likelihood estimator. This method makes it possible to use the information in physically close loci, whereas previous approaches could only use a set of independent loci to compute the likelihood.[13] Introduced a variation of the model that takes into account the effect ascertainment correction in the likelihood and MLE. This method is useful for data from loci that were selected because of their observed allele-frequencies (ie, ascertained). This method modifies the pruning algorithm of[6] so that the likelihood is corrected for ascertainment bias. In the same year[13] introduced a computational method for incorporating the effect of mutation at each branch in the pruning algorithm of.[6] This method makes it possible to apply the pruning algorithm to data from different species. This is because although the effect of new mutations can be ignored in closely related populations, their effect s are large when comparing different species. Finally,[9] has shown this model to be identifiable. The identifiability is an important desirable property of a statistical model; unidentifiable model parameters are ill-defined and therefore inference on such models could produce erroneous, confusing, and self-contradictory results.

Although the MLE of the divergence time is computed based on[4,6] model by the previous authors, a formula for the range of divergence time has not been described. The computation of the MLE requires computation of the likelihood and then maximization of the likelihood over a period of parameter values. Computation of a range or confidence interval for divergence time requires estimation of variance and MLE, which is a more complicated procedure.[4] computed an estimate for the variance through simulations, which could be used to compute a confidence interval. To estimate variance in this manner, one needs to first estimate the divergence time; then one needs to simulate the whole dataset a large number of times (say $M = 100$ times) using the estimated value of the divergence time as the real divergence time. Then, the divergence time is estimated from each of those $M = 100$ simulated datasets. The sample variance among the M estimated divergence time is taken as the estimated variance. However, this method takes a long time to compute, as one needs to simulate and estimate the divergence time $M = 100$ times (or a predetermined large number of times). A resampling approach (eg, bootstrap) would have a similar large time requirement.

Here, we present a formula for computing an asymptotic approximation for divergence time range. Our formula is based on asymptotic approximation of the variance of the MLE. In statistical literature, this approximation is known to be a first converging approximation.[14] Typically, the number of independent data-points (independent loci) is very high ($>10,000$) in genetic data and therefore this will be a close approximation. In addition, we also provide a formula for first and second order mixed derivatives of the likelihood and log-likelihood. This formula is useful as it can be used for maximizing the likelihood with Newton Raphson Method. To evaluate the performance of our method, we also estimated the coverage probability of the confidence interval through simulations, and established that the approximation is indeed quite good, and as accurate as a purely simulation-based approach. However, as demonstrated in the article, our method is much faster and could be computed in a fraction of time it takes to estimate the range using a simulation-based approach.

For demonstration purposes, we have used our method for estimating the range of the divergence time between two populations in HapMap data.[15] Our estimates are found to be of the same range as a recently published estimate of the divergence time of the same two populations.

Model and Definitions

In this section, we will briefly describe the model of.[4,6] This model assumes that the divergence time is sufficiently small so that we can ignore the effect of mutation in the site-frequency-spectrum after divergence. Consider populations A and B with MRCA O (Fig. 1). Let us assume that each locus exhibits two alleles, arbitrarily named '0' and '1', by 'allele count' we will mean the count of allele '1'. The model of[4,6] gives the probability distribution of the allele counts (r_A, r_B) from (haploid) sample of sizes (n_A, n_B) respectively. (Note that by "haploid" we mean a sampling unit. The data can be from either haploid or diploid organisms. A haploid sampling unit is as a set of chromosomes containing one chromosome of each type. Each diploid individual has two sampling units).

The parameters of this model are $\tau\,(>0)$, the divergence time in generations in the unit of effective population size, or the population scaled divergence time ($\tau = 2\,N_e\,t$), and $\theta(>0)$, a mutation parameter or a population scaled mutation rate ($\theta = 4\,N_e\,\mu$) where t, N_e and μ are, respectively, the number of generations since the divergence, the effective population size and (raw) mutation rate at the time of divergence. Note that we assume a molecular clock; ie, the scaled time between A and O is same as the scaled time between B and O. However, this assumption is not binding, and we will discuss relaxing this assumption later in this article.

Next, we will describe how the probability distribution of (r_A, r_B) is computed. First, we follow n_A and n_B lineages at populations A and B respectively to the MRCA O, and compute the probability distribution of the number of coalescent events k_A and k_B, respectively. Let n_{0A} and n_{0B} be the number of lineages at O that are ancestral to the sampled lineages at A and B, respectively, and let r_{0A} and r_{0B}, respectively, be the allele count out of them (Fig. 1). Note that $n_{0A} = n_A - k_A$ and $n_{0B} = n_B - k_B$. The distributions of n_0A and n_{0B} can be computed from the following formula (first proposed by).[16]

$$\Pr(n_{om} = i' \mid n_m = i;\tau) = \left(\prod_{j''=i'+1}^{i} \lambda_{j''}\right) \sum_{j=i'+1}^{i} \frac{e^{-\lambda_j \tau}}{\prod_{j'=i',j'\neq j}^{i}(\lambda_{j'} - \lambda_j)}, \quad (1)$$
$$= \sum_{j=i'}^{i} c_{ii'j}^{(1)} e^{-\lambda_j \tau},$$

$m = A, B$, where $\lambda_j = j(j-1)/2$ (1) and

$$c_{ii'j}^{(1)} = \frac{\left(\prod_{j''=i'+1}^{i} \lambda_{j''}\right)}{\prod_{j'=i',j'\neq j}^{i}(\lambda_{j'} - \lambda_j)}.$$

Let n_0 be the total number of lineages at O that are ancestral to the all lineages sampled at A and B. Using the fact $n_0 = n_{0A} + n_{0B}$ along with Eq. (1), one can compute the probability distribution of n_0. Let r_0 be (random) allele count out of the n0 lineages at the MRCA (Fig. 1). ($r_0 = r_{0A} + r_{0B}$.) The probability of r_0 given n_0 is given by a 'root distribution' that varies in different versions of the model. We use the root distribution used in[6] (symmetric beta-binomial)

$$\Pr(r_0 = j_0 \mid n_0 = i_0;\theta) = \binom{i_0}{j_0} \frac{\beta(j_0 + \theta, i_0 - j_0 + \theta)}{\beta(\theta,\theta)} \quad (2)$$

where $\beta(.,.)$ is the beta function; θ is the aforementioned mutation parameter which is to be estimated as well. Note that as mentioned in,[6] beta-binomial distribution model is due to the fact that the alleles at the root are binomial draws from the allele frequencies, and the allele frequency spectrum is modeled with a beta distribution (see, for example).[10]

Given n_{0A}, n_{0B}, and r_0, the distribution of (r_{0A}, r_{0B}) can be computed as

$$\Pr(r_{0A} = j_{0A}, r_{0B} = j_{0B} \mid r_0 = j_0, n_{0A} = i_{0A}, n_{0B} = i_{0B};\theta)$$
$$= \frac{\binom{j_{oA} + j_{0B}}{j_{0A}} \binom{i_{0A} + i_{0B} - j_{0A} - j_{0B}}{i_{0A} - j_{0A}}}{\binom{i_{0A} + i_{0B}}{i_{0A}}} \quad (3)$$
$$= c_{i_{oA} i_{0B} j_{0A} j_{0B}}^{(2)} \quad \text{(say)}$$

[4,6]. Then, given n_{0A}, n_{0B}, r_{0A} and r_{0B} the distribution of (r_A, r_B) can be computed as

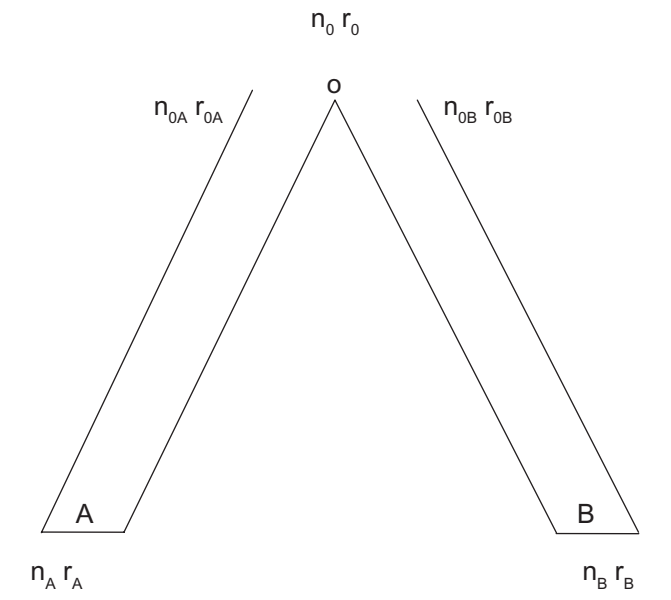

Figure 1. Variables associated with our model.

$$\Pr\left(r_m = j_m \mid r_{0m} = j_{0m}, n_{0m} = i_{0m}; n_m = i_m\right) = \begin{cases} \dfrac{\beta(j_m, i_m - j_m)}{\beta(j_{0m}, i_{0m} - j_{0m})}\binom{i_m - i_{0m}}{j_m - j_{0m}}, & 0 < j_m < i_m \text{ and } 0 < j_{0m} < i_{0m}, \\ 1, & 0 = j_m = j_{0m} \text{ or } 0 = i_m - j_m = i_{0m} - j_{0m}, \\ 0, & \text{otherwise} \end{cases} = c^{(3)}_{i_m i_{0m} j_{0m} j_m} \quad \text{(say)} \tag{4}$$

$m = A, B$.[6]

Next we combine the Eqs. (1, 2, 3, 4). With the observed allele counts are (r_A, r_B), the likelihood of (τ, θ) can be computed as

$$\begin{aligned} L(\tau,\theta) &= L\left(\tau,\theta;(r_A,r_B) = (j_A,j_B)\right) \\ &= \Pr((r_A,r_B) = (j_A,j_B);(n_A,n_B) = (i_A,i_B),\tau,\theta) \\ &= \sum_{i_{0A}=1}^{i_A}\sum_{i_{0B}=1}^{i_B} \Pr(n_{0A} = i_{0A}; n_A = i_A,\tau)\Pr(n_{0B} = i_{0B}; n_B = i_B,\tau) \\ &\quad \times \sum_{j_{0A}=0}^{i_{0A}}\sum_{j_{0B}=0}^{i_{0B}} \Pr(r_0 = j_{0A} + j_{0B} \mid n_0 = i_{0A} + i_{0B};\theta) \\ &\quad \times \Pr(r_{0A} = j_{0A}, r_{0B} = j_{0B} \mid r_0 = j_{0A} + j_{0B}, n_{0A} = i_{0A}, n_{0B} = i_{0B}) \\ &\quad \times \Pr(r_A = j_A \mid r_{0A} = j_{0A}, n_{0A} = i_{0A}; n_A = r_A)\Pr(r_B = j_B \mid r_{0B} = j_{0B}, n_{0B} = i_{0B}; n_B = r_B) \\ &= \sum_{i_{0A}=1}^{i_A}\sum_{i_{0B}=1}^{i_B}\left(\sum_{i_1=i_{0A}}^{i_A}\sum_{i_2=i_{0B}}^{i_B} c^{(1)}_{i_A i_{0A} i_1} c^{(1)}_{i_B i_{0B} i_2} e^{-(\lambda_{i_1}+\lambda_{i_2})\tau}\right) \\ &\quad \times \sum_{j_{0A}=0}^{i_{0A}}\sum_{j_{0B}=0}^{i_{0B}}\binom{i_{0A}+i_{0B}}{j_{0A}+j_{0B}}\frac{\beta(j_{0A}+j_{0B}+\theta, i_{0A}+i_{0B}-j_{0A}-j_{0B}+\theta)}{\beta(\theta,\theta)} \\ &\quad \times c^{(2)}_{i_{0A} i_{0B} j_{0A} j_{0B}} c^{(3)}_{i_A i_{0A} j_{0A} j_A} c^{(3)}_{i_B i_{0B} j_{0B} j_B} \\ &= \sum_{i_1=1}^{i_A}\sum_{i_2=1}^{i_B} e^{-(\lambda_{i_1}+\lambda_{i_2})\tau}\sum_{i_{0A}=1}^{i_1}\sum_{i_{0B}=1}^{i_2} c^{(1)}_{i_A i_{0A} i_1} c^{(1)}_{i_B i_{0B} i_2} \\ &\quad \times \sum_{j_{0A}=0}^{i_{0A}}\sum_{j_{0B}=0}^{i_{0B}}\binom{i_{0A}+i_{0B}}{j_{0A}+j_{0B}}\frac{\beta(j_{0A}+j_{0B}+\theta, i_{0A}+i_{0B}-j_{0A}-j_{0B}+\theta)}{\beta(\theta,\theta)} \\ &\quad \times c^{(2)}_{i_{0A} i_{0B} j_{0A} j_{0B}} c^{(3)}_{i_A i_{0A} j_{0A} j_A} c^{(3)}_{i_B i_{0B} j_{0B} j_B}. \end{aligned} \tag{5}$$

Note that setting $\tau = 0$ does not maximize the above expression, as the coeffcients of the exponential terms could be positive or negative.

Thus, we have described the full model. Maximum likelihood estimate (MLE) of the divergence time is computed by numerically maximizing right side of Eq. (5) above. In the next section we will present an estimator of the divergence time range, rather than a point-estimator of divergence time.

Methods

Let the MLE of (τ, θ) be $(\hat{\tau}_{\text{MLE}}, \hat{\theta}_{\text{MLE}})$. Using the standard statistical results,

$$\sqrt{L}\left(\begin{pmatrix}\hat{\tau}_{\text{MLE}}\\ \hat{\theta}_{\text{MLE}}\end{pmatrix} - \begin{pmatrix}\tau\\ \theta\end{pmatrix}\right) \to_d \text{Normal}_2(0, \text{Inf}^{-1}(\tau,\theta)) \tag{6}$$

where $\rightarrow_d$ denotes convergence in distribution (see, for example);[7] L is the number of independent data-points (independent loci in our case), Normal_2 denotes a bivariate normal distribution, and Inf denotes Fisher's Information matrix:

$$\text{Inf}\,(\tau,\theta)=E\left[-\begin{pmatrix}\dfrac{\partial^2}{\partial\tau^2}\log_e L(\tau,\theta) & \dfrac{\partial^2}{\partial\tau\partial\theta}\log_e L(\tau,\theta)\\ \dfrac{\partial^2}{\partial\tau\partial\theta}\log_e L(\tau,\theta) & \dfrac{\partial^2}{\partial\theta}\log_e L(\tau,\theta)\end{pmatrix}\right] \tag{7}$$

where $L(\tau, \theta) = L(\tau, \theta; (r_A, r_B) = (j_A, j_B))$ is the likelihood. Using Eq. (6), one can estimate a $(1 - \alpha)$ confidence interval for τ as

$$\{\hat{\tau}_{\text{MLE}} - z_{1-\alpha/2}\sqrt{\text{Inf}^{(11)}(\hat{\tau}_{\text{MLE}},\hat{\theta}_{\text{MLE}})/L},\ \hat{\tau}_{\text{MLE}} + z_{1-\alpha/2}\sqrt{\text{Inf}^{(11)}(\hat{\tau}_{\text{MLE}},\hat{\theta}_{\text{MLE}})/L}\} \tag{8}$$

and a $(1 - \alpha)$ confidence interval for θ as

$$\{\hat{\theta}_{\text{MLE}} - z_{1-\alpha/2}\sqrt{\text{Inf}^{(22)}(\hat{\tau}_{\text{MLE}},\hat{\theta}_{\text{MLE}})/L},\ \hat{\theta}_{\text{MLE}} + z_{1-\alpha/2}\sqrt{\text{Inf}^{(22)}(\hat{\tau}_{\text{MLE}},\hat{\theta}_{\text{MLE}})/L}\}, \tag{9}$$

where $\text{Inf}^{(ij)}$ is the element (i, j) of Inf^{-1}, $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of standard normal distribution. Next, we obtain a simpler expression for Eqs. (8, 9).

Using Lemma 1 in the Appendix A (Eqs. (13, 14)) and Eq. (5) it follows that

$$\begin{aligned}
&\frac{\partial^{l+m}}{\partial\tau^l\partial\theta^m}L(\tau,\theta;(r_A,r_B)=(j_A,j_B))\\
&=\sum_{i_1=1}^{i_A}\sum_{i_2=1}^{i_B}(-(\lambda_{i1}+\lambda_{i2}))^l e^{-((\lambda_{i1}+\lambda_{i2})\tau}\sum_{i_{0A}=1}^{i_1}\sum_{i_{0B}=1}^{i_2}c^{(1)}_{i_Ai_{0A}i_1}c^{(1)}_{i_Bi_{0B}i_2}\\
&\quad\times\sum_{j_{0A}=0}^{i_{0A}}\sum_{j_{0B}=0}^{i_{0B}}\binom{i_{0A}+i_{0B}}{j_{0A}+j_{0B}}\frac{\beta(j_{0A}+j_{0B}+\theta,i_{0A}+i_{0B}-j_{0A}-j_{0B}+\theta)}{\beta(\theta,\theta)}\\
&\quad\times\left(\delta^m(\theta,j_{0A}+j_{0B},i_{0A}+i_{0B}-j_{0A}-j_{0B})+1_{\{m=2\}}\delta_1(\theta,j_{0A}+j_{0B},i_{0A}+i_{0B}-j_{0A}-j_{0B})\right)\\
&\quad\times c^{(2)}_{i_{0A}i_{0B}j_{0A}j_{0B}}c^{(3)}_{i_Ai_{0A}j_{0A}j_A}c^{(3)}_{i_Bi_{0B}j_{0B}j_B}
\end{aligned} \tag{10}$$

for $l, m \in \{1, 2\}$. Also, note that

$$\begin{aligned}
&\frac{\partial^{l+m}}{\partial\tau^l\partial\theta^m}\log_e L(\tau,\theta;(r_A,r_B)=(j_A,j_B))\\
&=\frac{\dfrac{\partial^{l+m}}{\partial\tau^l\partial\theta^m}L(\tau,\theta;(r_A,r_B)=(j_A,j_B))}{L(\tau,\theta;(r_A,r_B)=(j_A,j_B))}\\
&\quad-\frac{\left(\dfrac{\partial}{\partial\tau}L(\tau,\theta;(r_A,r_B)=(j_A,j_B))\right)^l\left(\dfrac{\partial}{\partial\theta}L(\tau,\theta;(r_A,r_B)=(j_A,j_B))\right)^m}{\left(L(\tau,\theta;(r_A,r_B)=(j_A,j_B))\right)^2}
\end{aligned} \tag{11}$$

$(l, m) \in \{(1, 1), (2, 0), (0, 2)\}$ and hence,

$$
\begin{aligned}
&E\left[\frac{\partial^{l+m}}{\partial\tau^{l}\partial\theta^{m}}\log_{e} L(\tau,\theta;(r_A,r_B)=(j_A,j_B))\right]\\
&=\sum_{j_A=0}^{n_A}\sum_{j_B=0}^{n_B}\left[\frac{\partial^{l+m}}{\partial\tau^{l}\partial\theta^{m}}L(\tau,\theta;(r_A,r_B)=(j_A,j_B))\right]\\
&-\sum_{j_A=0}^{n_A}\sum_{j_B=0}^{n_B}\left[\frac{\left(\frac{\partial}{\partial\tau}L(\tau,\theta;(r_A,r_B)=(j_A,j_B))\right)^{l}\left(\frac{\partial}{\partial\theta}L(\tau,\theta;(r_A,r_B)=(j_A,j_B))\right)^{m}}{L(\tau,\theta;(r_A,r_B)=(j_A,j_B))}\right]
\end{aligned}
\tag{12}
$$

$(l, m) \in \{(1, 1), (2, 0), (0, 2)\}$.

Estimation. Using Eqs. (5, 7–12), one can numerically compute confidence intervals of τ and θ as follows.

First, $\hat{\tau}_{\text{MLE}}$ and $\hat{\theta}_{\text{MLE}}$ are computed maximizing the likelihood given in Eq. (5). The maximization may be done over a grid of values of τ and θ; alternatively, the Newton Raphson Method may be used; the derivatives of likelihood or log-likelihood for use in Newton Raphson Method can be obtained by using the expressions of derivatives in the likelihood in Eq. (10) in the expression of Eq. (11).

Once we have $\hat{\tau}_{\text{MLE}}$ and $\hat{\theta}_{\text{MLE}}$, we can compute $\hat{\tau}_{\text{MLE}},\hat{\theta}_{\text{MLE}}$ using Eqs. (10,12) with (τ, θ) substituted by $\hat{\tau}_{\text{MLE}},\hat{\theta}_{\text{MLE}}$. Next,is $\text{Inf}(\hat{\tau}_{\text{MLE}},\hat{\theta}_{\text{MLE}})^{-1}$ numerically computed numerically and subsequently the confidence intervals of Eqs. (8,9) are computed.

Results: Simulation and Comparison with Direct Simulation Method

We have simulated $N = 1,000$ datasets for each combination of sample sizes $n = n_A = n_B$, divergence time τ and number of independent loci L; n can take values 4 and 8; τ can take values 0.01, 0.02, 0.05, 0.075, 0.1, 0.15 and 0.2; L can take values 10,000 and 100,000. The mutation parameter was kept fixed at $\theta = 4 \times 10{,}000 \times (1.1 \times 10^{-8})$ to reflect a (human) effective population size of 10, 000 and a mutation rate of 1.1×10^{-8}.

Mechanism of model generation. Following the model of[4,6] we started with a divergence time τ, mutation parameter θ, sample size (identical for the two populations) $n = n_A = n_B$, and a predetermined number of independent loci L. For each locus the process is identical and independent (given the parameters). Therefore, we will describe the process for a single locus only.

For a given locus, we simulate the model (described in Section 2) as follows. First n_{0A} and n_{0B} are simulated from n_A and n_B using Eq. (1). Next, n_0 is simulated (or computed) as $n_0 = n_{0A}+n_{0B}$. Then, r_0 is simulated from n_0 and θ from symmetric beta-binomial distribution (Eq. (2)) as in.[6] Symmetric beta-binomial distribution is used in[6] to characterize the allele-frequency spectrum. That is, it is assumed in[6] that the allele-frequency spectrum over the L loci has a symmetric Beta distribution, and r_0 given n_0 is a randomly drawn allele-count from a sample of n haploids where the allele frequency is P (which is a random draw from the allele-frequency spectrum and has a symmetric beta distribution with parameter θ). Thus, r_0 given n_0 has a symmetric beta-binomial distribution with parameter θ. Next, we simulate r_{0A} and r_{0B} from n_0, r_0, n_{0A} and n_{0B} using Eq. (3). Then r_m is simulated from n_m, n_{0m} and r_{0m} using Eq. (4) for $m = A$ and B. This process is repeated independently (given the parameters) L times to generate the allele counts r_A and r_B for each of the L loci.

For each combination of (n, τ, L), an MLE $(\hat{\tau}_{\text{MLE}},\hat{\theta}_{\text{MLE}})$ was computed by maximizing the full likelihood over τ and θ for each $N = 1,000$ repetitions. Thus, for each combination of (n, τ, L), we have $N = 1,000$ estimates of $(\hat{\tau}_{\text{MLE}},\hat{\theta}_{\text{MLE}})$. Then, applying our methods on these estimates, 95% confidence intervals for τ was estimated for each of $N = 1,000$ repetitions for each combination of (n, τ, L). Then for each combination of (n, τ, L), we computed the average of the confidence intervals over $N = 1,000$ estimated values (reported in Table 1 at the CI_{Asymp} columns).

Next, for each combination of (n, τ, L), we have also estimated the probability of the true τ falling into the estimated confidence interval (coverage probability) as the total number of time the true value was in the estimated interval (among $N = 1,000$ repetitions) divided by N. The results are in Table 1 at the CI_{Asymp} columns.

Next, we compared the performance of our method with the method of computing the confidence interval using simulation estimation of variance[4] using the same simulated data. We have briefly described their method in Section 1. As we had done for our method, we estimated confidence interval for each of $N = 1,000$ repetitions and for each combination of (n, τ, L). (Note that, this involved resimulating $M = 100$ datasets for each estimated $\hat{\tau}_{\text{MLE}}$ and estimating τ from the resimulated dataset, and thus a total of $N \times M = 10^5$ simulations and estimations.) Then, we have computed the average estimated 95% confidence interval and coverage probability from $N = 1,000$ simulated datasets for each combination of (n, τ, L). That is, after estimating the confidence intervals for each dataset, we computed the average of the confidence intervals over $N = 1,000$ estimated values for each combination of (n, τ, L) (reported in Table 1). The coverage probabilities are estimated as before as the total number of time the true value was in

Table 1: Simulation: confidence interval for τ and estimated coverage probability P.

	$L = 10{,}000$ CI_{ASYMP}	SIM. VAR. CI	$L = 100{,}000$ CI_{ASYMP}	SIM. VAR. CI
$\tau = 0.01$ $n = 4$	CI: {0.01 ± 0.0010} $P = 0.951$	{0.01 ± 0.0009} $P = 0.955$	CI: {0.01 ± 0.0003} $P = 0.949$	{0.01 ± 0.0003} $P = 0.955$
$\tau = 0.01$ $n = 8$	CI: {0.01 ± 0.0004} $P = 0.947$	{0.01 ± 0.0005} $P = 0.954$	CI: {0.01 ± 0.0001} $P = 0.954$	{0.01 ± 0.0001} $P = 0.945$
$\tau = 0.02$ $n = 4$	CI: {0.02 ± 0.0015} $P = 0.945$	{0.02 ± 0.0014} $P = 0.960$	CI: {0.02 ± 0.0005} $P = 0.950$	{0.02 ± 0.0003} $P = 0.953$
$\tau = 0.02$ $n = 8$	CI: {0.02 ± 0.0006} $P = 0.948$	{0.02 ± 0.0005} $P = 0.948$	CI: {0.02 ± 0.0002} $P = 0.942$	{0.02 ± 0.0002} $P = 0.944$
$\tau = 0.05$ $n = 4$	CI: {0.05 ± 0.0025} $P = 0.954$	{0.05 ± 0.0029} $P = 0.959$	CI: {0.05 ± 0.0008} $P = 0.955$	{0.05 ± 0.0010} $P = 0.953$
$\tau = 0.05$ $n = 8$	CI: {0.05 ± 0.0012} $P = 0.951$	{0.05 ± 0.0010} $P = 0.944$	CI: {0.05 ± 0.0004} $P = 0.951$	{0.05 ± 0.0005} $P = 0.945$
$\tau = 0.075$ $n = 4$	CI: {0.075 ± 0.0032} $P = 0.954$	{0.075 ± 0.0036} $P = 0.949$	CI: {0.075 ± 0.0010} $P = 0.942$	{0.075 ± 0.0011} $P = 0.950$
$\tau = 0.075$ $n = 8$	CI: {0.075 ± 0.0016} $P = 0.944$	{0.075 ± 0.0015} $P = 0.943$	CI: {0.075 ± 0.0005} $P = 0.952$	{0.075 ± 0.0005} $P = 0.961$
$\tau = 0.1$ $n = 4$	CI: {0.10 ± 0.0040} $P = 0.964$	{0.10 ± 0.0041} $P = 0.945$	CI: {0.10 ± 0.0013} $P = 0.956$	{0.10 ± 0.0011} $P = 0.961$
$\tau = 0.1$ $n = 8$	CI: {0.10 ± 0.0022} $P = 0.949$	{0.10 ± 0.0018} $P = 0.950$	CI: {0.10 ± 0.0007} $P = 0.949$	{0.10 ± 0.0007} $P = 0.948$
$\tau = 0.15$ $n = 4$	CI: {0.15 ± 0.0055} $P = 0.957$	{0.15 ± 0.0049} $P = 0.954$	CI: {0.15 ± 0.0017} $P = 0.948$	{0.15 ± 0.0016} $P = 0.955$
$\tau = 0.15$ $n = 8$	CI: {0.15 ± 0.0034} $P = 0.951$	{0.15 ± 0.0038} $P = 0.949$	CI: {0.15 ± 0.0011} $P = 0.957$	{0.15 ± 0.0011} $P = 0.956$
$\tau = 0.2$ $n = 4$	CI: {0.2 ± 0.0071} $P = 0.940$	{0.2 ± 0.0076} $P = 0.949$	CI: {0.2 ± 0.0022} $P = 0.954$	{0.2 ± 0.0024} $P = 0.957$
$\tau = 0.2$ $n = 8$	CI: {0.2 ± 0.0050} $P = 0.944$	{0.2 ± 0.0054} $P = 0.951$	CI: {0.2 ± 0.0016} $P = 0.957$	{0.2 ± 0.0019} $P = 0.948$

the estimated interval divided by N. The results are shown in Table 1.

For both CI_{Asymp} and simulated CI the estimated ranges have small lengths. As expected, the length becomes smaller with larger n and larger L. For $\tau = 0.01$ the ranges have radii 0.0010 (for $n = 4$, $L = 10,000$, CI_{Asymp}) to 0.0001 (for $n = 8$, $L = 100,000$, both methods). The length of the radii increases with τ. For $\tau = 0.02$ the ranges have radii 0.0015 (for $n = 4$, $L = 10,000$, CI_{Asymp}) to 0.0002 (for $n = 8$, $L = 100,000$, both methods). For $\tau = 0.05$ the ranges have radii 0.0029 (for $n = 4$, $L = 10,000$, simulated CI) to 0.0004 (for $n = 8$, $L = 100,000$, CI_{Asymp}). The radius increases up with an approximate proportionality with the true value of τ. For $\tau = 0.1$ and 0.2 the ranges have radii 0.0041 and 0.0076, respectively, (for $n = 4$, $L = 10,000$ simulated CI) to 0.0007 (for $n = 8$, $L = 100,000$ both methods) and 0.0016 (for $n = 8$, $L = 100,000$ CIASYMP), respectively.

A comparison of the two methods reveals no significant difference in the coverage probabilities of the lengths of the intervals. In Wilcoxon signed-rank tests we found no significant difference between coverage probabilities of the two methods ($P > 0.3$) as well as between lengths of the intervals ($P > 0.9$). However, computing time was an order of magnitude faster for CI_{Asymp} than for the simulated CI method. Once we have $\hat{\tau}_{MLE}$ and $\hat{\theta}_{MLE}$, it takes less than 30 minutes to compute the CI using CI_{Asymp} using a R v3.0.2013-05-12[17] code in a 2.54 GHz Dual Core processor. Using the same computer and same version of R, simulated CI (with at least $M = 100$ simulations for estimating the variance) takes more than a day to compute. This is because each simulation involves simulating L independent loci and maximization of the multi-locus likelihood over τ and θ using the allele-counts from these loci.

Results: Applications to HapMap Data

For the purpose of demonstrating our method and comparing its performance with known results, we applied our method to a subset of HapMap data.[15] Specifically, we estimated a confidence interval for the divergence time between HCB (Han Chinese from Be-jing, China) and CEU (United States residents of northern and western European ancestry) populations using 112 independent SNP loci in Chromosome 19. To reduce the computational load, we only considered a random sub-sample of 8 *unrelated* haploids from each population.

Our data consist of the allele-count in 8 individuals in each population for 112 SNP loci. The 112 SNPs were selected at least 0.5 MB apart from each other to ensure the independence of the coalescent trees in the lineages of 8 haploids.

The estimated scaled divergence time was $\hat{\tau} = 0.16$, and the estimated 95% confidence intervals was. [0.125, 0.195]. These numbers, when transformed in years using an overall effective population size of 3,100 (see, for example)[18] for HCB and CEU populations and generation time of 25 years, produce an estimated range of between 19,375–30,225 years ago. If an overall effective population size of 4,000 is assumed, then this produces as estimated range of between 25,000–39,000 years ago. Note that these ranges roughly match with other recent estimates of divergence time between HCB and CEU (see, for example).[19]

Discussion

We have presented a formula for computing divergence time range using the MLE on a coalescent model. As MLE is asymptotically effcient, our estimated confidence interval has a high degree of accuracy. We have also presented formulae for first and second order mixed derivatives of likelihood and log-likelihood, which is useful for computation of the MLE.

Our simulation study shows that our method produces small confidence intervals with appropriate coverage of 95%. Thus, it reduces the amount of uncertainty regarding the actual divergence time. Although MLE of divergence time has been computed before using our model, an expression for the range of divergence time has never been derived because of the complexity of the expression. By deriving this expression we made it possible for the range of the divergence time to be estimated, and by evaluating its performance we established its usefulness.

The radii of the confidence intervals and the coverage probabilities produced by our methods have been shown to be statistically equivalent to that of the simulation-based estimator. However, we found that our method is an order of magnitude faster. This is expected because in a simulation or resampling-based method the data needs to be simulated a large number of times, and then confidence intervals needs to be reestimated for each resimulation or resample.

Certain assumptions are made about the underlying model. The accuracy of the esti-mated variance and range may be affected if the data do not fit the underlying model. For example, a molecular clock is assumed. That is, the scaled time τ between A and O is assumed to be same as that between B and O. However, if the effective population sizes in populations a and b are very different then this will induce a bias in the estimated variance, as well as the estimated range. However, it is straightforward to extend our method for models without molecular clock. If a molecular clock is not assumed then one needs to separately estimate scaled divergence times τ_A and τ_B for Populations a and b. It is straightforward, albeit tedious, to modify Eqs. (6–16) for parameters (τ_A, τ_B, θ), rather than (τ,θ). Thus, analogous expressions for the variance and the range of (τ_A, τ_B, θ) can be computed if a molecular clock is not assumed.

Another departure from our model would be from the assumption of biallelic loci. Following the same principle of tracking coalescing lineages back to the MRCA, and then following the allele-types to the present time, one can modify Eqs. 1–5 for more than two alleles. However, for more than two alleles, one needs to replace the symmetric beta-binomial distribution at the MRCA (which arises from assuming a symmetric beta distribution for biallelic allele-frequency spectrum) with its "multiple-count" version: a symmetric Dirichlet-Multinomial distribution (which arises from assuming a symmetric Dirichlet distribution for multiallelic allele-frequency spectrum). Once the likelihood is computed using modified version of Eqs. 1–5, then Eqs. 6–16 can be modified for estimating the variance and the range from the multiallelic likelihood.

We assumed that the effect of mutation between the point of divergence and the present is negligible. This is an appropriate assumption if the two populations are closely related (and consequently the divergence time is small). For large divergence times this assumption is not appropriate, and the effect of mutation may create an amount of difference between the two population that is higher than expected from an ignore-mutation model". As a result, an erroneously larger divergence time may be estimated, which will also induce an upward bias in the estimated variance and the estimated range.

previous studies have suggested that this model is appropriate for populations within the same species.[5–8] This designation is intended to serve as an upper bound for the divergence time to be modeled. In the absence of a mathematically concrete upper bound for divergence time for this model, we also use this convention. Thus, although we do not have a more concrete upper bound, we suggest using our methods for doing inference on divergence time between populations within the same species. (A lower bound is not necessary as the model fits well for small divergence times as the amount of mutation will be very small in such cases.) Moreover, a recent article[13] introduces a version of the[4,6] model that takes into account the effect of mutation. An appropriate extension of our methods to the[13] model needs to be derived for estimating the variance and the range of larger divergence times. With such an extension, our methods could be used to estimate variance and range of species divergence times.

A possible disadvantage of our method would be in a scenario where significant amount of migration has taken place between the two populations after the divergence. In the presence of migration, there will be less variation than expected in a "no migration" scenario. Consequently, divergence time will be underestimated, which will also induce a downward bias in the estimated variance and the estimated range. Another limitation is that our method uses asymptotic approximation. Thus, our method is only applicable when we have a large

number of independent loci. To quantify a large number of loci, a rule of thumb used by statisticians is that the asymptotic methods are to be used if there are at least 30 independent data-points. Thus, as long as we have allele counts for at least 30 independent loci, our method can be used. As most modern datasets have more than 30 independent loci, this is not a real limitation. Note that, as the effect of mutation after divergence is assumed to be negligible and is ignored, the difference in mutation rate between the two populations does not play a part in our method.

We applied our method to estimate a range for the divergence time between CEU and HCB populations from the HapMap data. Our estimates are found to be in the same range as a recently estimated divergence time between these two populations.[19]

A possible extension of this method could be computation of the confidence interval of the branch-length of a phylogenetic tree given the tree-topology. This would require creating an algorithm that could effciently and systematically compute the derivatives of the likelihood of the tree. Another possible extension could be using dependent loci to estimate the range of divergence time. This will have potential applications in high-resolution NextGen Sequencing data.

Acknowledgment

The author is grateful for constructive comments by anonymous reviewers.

Author Contributions

Conceived and designed the experiments: AR. Analyzed the data: AR. Wrote the first draft of the manuscript: AR. Contributed to the writing of the manuscript: AR. Agree with manuscript results and conclusions: AR. Jointly developed the structure and arguments for the paper: AR. Made critical revisions and approved final version: AR. The author reviewed and approved of the final manuscript.

DISCLOSURES AND ETHICS

As a requirement of publication the authors have provided signed confirmation of their compliance with ethical and legal obligations including but not limited to compliance with ICMJE authorship and competing interests guidelines, that the article is neither under consideration for publication nor published elsewhere, of their compliance with legal and ethical guidelines concerning human and animal research participants (if applicable), and that permission has been obtained for reproduction of any copyrighted material. This article was subject to blind, independent, expert peer review. The reviewers reported no competing interests.

REFERENCES

1. Rona LDP, Carvalho-Pinto CJ, Mazzoni CJ, Peixoto AA. Estimation of divergence time between two sibling species of the Anopheles (Kerteszia) cruzii complex using a multilocus approach. *BMC Evolutionary Biology*. 10, 2010. DOI:10.1186/1471-2148-10-91.
2. Wang X, Gowik U, Tang H, Bowers JE, Westhoff P, Paterson AH. Comparative genomic analysis of c4 photosynthetic pathway evolution in grasses. *Genome Biology*. 2009;10:R68.
3. Langergraber KE, Prfer K, Rowney C, et al. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *PNAS*. 2012;109:15716–21.
4. Nielsen R, Mountain JL, Huelsenbeck JP, Slatkin M. Maximum likelihood estimation of population divergence times and population phylogeny in models without mutation. *Evolution*. 1998;52:669–77.
5. RoyChoudhury A. *Likelihood inference for population structure, using the coalescent*. PhD thesis, University of Washington, 2006.
6. RoyChoudhury A, Felsenstein J, Thompson EA. A two-stage pruning algorithm for likelihood computation for a population tree. *Genetics*. 2008;180:1095–105.
7. RoyChoudhury A. Composite likelihood-based inferences on genetic data from depen-dent loci. *Journal of Mathematical Biology*. 62:65–80, 2011.
8. RoyChoudhury A, Thompson EA. Ascertainment correction for a population tree via a pruning algorithm for likelihood computation. *Theoretical Population Biology*. 2012;82:59–65.
9. RoyChoudhury A. Identifiability of a coalescent-based population tree model. arXiv, pages *arXiv*:1304.3691 [q–bio.PE], 2013.
10. Ewens WJ. *Mathematical Population Genetics*. Springer, 2004.
11. Degnan JH, Rosenberg NA. Gene tree discordance, phylogenetic inference, and the multispecies coalescent. *Trends in Ecology and Evolution*. 2009;24:332–40.
12. Nielsen R, Slatkin M. Likelihood analysis of ongoing gene flow and historical association. *Evolution*. 2000;54:44–50.
13. Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, RoyChoudhury A. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol Bol Evol*. 2012;29:1917–32.
14. Lehmann EL, Casella G. *Theory of Point Estimation*. New York: Springer, 1998.
15. The International HapMap Consortium. The International HapMap Project. *Nature*. 2003;426:789–96.
16. Takahata N, Nei M. Gene genealogy and variance of interpopulational nucleotide differences. *Genetics*. 1985;110:325–44.
17. R Core Team. R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.
18. Tenesa A, Navarro P, Hayes BJ, et al. Recent human effective population size estimated from linkage disequilibrium. *Genome Research*. 2007;17:520–6.
19. Gravel S, Henn BM, Gutenkunst RN, et al. The 1000 Genomes Project, and C. D. Bustamante. Demographic history and rare allele sharing among human populations. *PNAS*. 2011;108:11983–8.

A Lemma 1

Lemma 1:

$$\frac{\partial}{\partial\theta}\frac{\beta(x+\theta,y+\theta)}{\beta(\theta,\theta)}=\frac{\beta(x+\theta,y+\theta)}{\beta(\theta,\theta)}\delta(\theta,x,y), \tag{13}$$

$$\frac{\partial^2}{\partial\theta^2}\frac{\beta(x+\theta,y+\theta)}{\beta(\theta,\theta)}=\frac{\beta(x+\theta,y+\theta)}{\beta(\theta,\theta)}(\delta(\theta,x,y)^2+\delta_1(\theta,x,y)), \tag{14}$$

where

$$\delta(\theta,x,y)=\psi(\theta+x)+\psi(\theta+y)-2\psi(2\theta+x+y)-2\psi(\theta)+2\psi(2\theta)$$
$$\delta_1(\theta,x,y)=\psi_1(\theta+x)+\psi_1(\theta+y)-4\psi_1(2\theta+x+y)-2\psi_1(\theta)+4\psi_1(2\theta)$$

where $\psi(.)$ is the digamma function and $\psi_1(.)$ is the trigamma function.

Proof:

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\beta(x+\theta,y+\theta)\\
&=\frac{\partial}{\partial\theta}\frac{\Gamma(x+\theta)\Gamma(y+\theta)}{\Gamma(x+y+2\theta)}\\
&=\frac{\Gamma(x+y+2\theta)(\Gamma'(x+\theta)\Gamma(y+\theta)+(\Gamma(x+\theta)\Gamma'(y+\theta))-2\Gamma(x+\theta)\Gamma(y+\theta)\Gamma'(x+y+2\theta)}{\Gamma(x+y+2\theta)^2}\\
&=\frac{\Gamma(x+\theta)\Gamma(y+\theta)}{\Gamma(x+y+2\theta)}[\psi(x+\theta)+\psi(y+\theta)-2\psi(x+y+2\theta)]\\
&=\beta(x+\theta,y+\theta)[\psi(x+\theta)+\psi(y+\theta)-2\psi(x+y+2\theta)].
\end{aligned} \tag{15}$$

From Eq. (15) it also follows that

$$\frac{\partial}{\partial\theta}\beta(\theta,\theta)=\beta(\theta,\theta)[2\psi(\theta)-2\psi(\theta)]. \tag{16}$$

From Eqs. (15,16)

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\frac{\beta(x+\theta,y+\theta)}{\beta(\theta,\theta)}\\
&=\frac{\beta(\theta,\theta)\beta(x+\theta,y+\theta)[\psi(x+\theta)+\psi(y+\theta)-2\psi(x+y+2\theta)]}{\beta(\theta,\theta)^2}\\
&\quad-\frac{\beta(\theta,\theta)\beta(x+\theta,y+\theta)[2\psi(\theta)-2\psi(2\theta)]}{\beta(\theta,\theta)^2}\\
&=\frac{\beta(x+\theta,y+\theta)}{\beta(\theta,\theta)}\delta(\theta,x,y)
\end{aligned} \tag{16}$$

Thus, Eq. (13) is proven.

$$
\begin{aligned}
&\frac{\partial^2}{\partial\theta^2}\frac{\beta(x+\theta,y+\theta)}{\beta(\theta,\theta)}\\
&=\frac{\partial}{\partial\theta}\frac{\beta(x+\theta,y+\theta)}{\beta(\theta,\theta)}\delta(\theta,x,y)\\
&=\left[\frac{\partial}{\partial\theta}\frac{\beta(x+\theta,y+\theta)}{\beta(\theta,\theta)}\right]\delta(\theta,x,y)+\frac{\beta(x+\theta,y+\theta)}{\beta(\theta,\theta)}\frac{\partial}{\partial\theta}\delta(\theta,x,y)\\
&=\left[\frac{\beta(x+\theta,y+\theta)}{\beta(\theta,\theta)}\delta(\theta,x,y)\right]\delta(\theta,x,y)+\frac{\beta(x+\theta,y+\theta)}{\beta(\theta,\theta)}\frac{\partial}{\partial\theta}\delta(\theta,x,y)\\
&=\frac{\beta(x+\theta,y+\theta)}{\beta(\theta,\theta)}\left(\delta(\theta,x,y)^2+\frac{\partial}{\partial\theta}\delta(\theta,x,y)\right)
\end{aligned}
$$

Eq. (14) follows from the above equation after noting that

$$\frac{\partial}{\partial\theta}\delta(\theta,x,y)=\delta_1(\theta,x,y).$$

Thus, we have proven Lemma 1.