MOLECULAR ECOLOGY

Molecular Ecology (2011) **20**, 3956–3962 doi: 10.1111/j.1365-294X.2011.05247.x

NEWS AND VIEWS

COMMENT

On the nonidentifiability of migration time estimates in isolation with migration models

VITOR C. SOUSA, AUDE GRELAUD and JODY HEY

Department of Genetics, Rutgers the State University of New Jersey, 604 Allison Road, Piscataway, NJ 08854, USA

Keywords: coalescent, divergence population genetics, gene flow, migration, speciation

Received 18 May 2011; revision received 12 July 2011; accepted 21 July 2011

In recent years, many studies have found evidence of gene flow between diverging populations by analyzing genetic data under an Isolation with Migration (IM) model (Pinho and Hey 2010). Given evidence of gene exchange, investigators often then wish to inquire of the time when gene flow occurred (e.g. Won & Hey 2005; Becquet & Przeworski 2009). For example, a model of divergence with gene flow would be suggested whether gene flow occurred early or throughout the divergence process, whereas secondary contact would be the likely interpretation if gene flow was found to only have occurred after divergence had been ongoing for some time. Recently, Strasburg and Rieseberg (2011) assessed the quality of estimates for the time of migration events using the method currently implemented in the IMA2 program (Hey 2010). They found that the credible intervals of estimated times were so wide as to make the method unsuitable for the question. These results suggest that some conclusions of previous studies that draw upon the posterior distribution for times of migration should be discounted (e.g. Won & Hey 2005; Niemiller *et al.* 2008; Strasburg *et al.* 2008; Nadachowska & Babik 2009).

The Strasburg & Rieseberg (2011) study reports results from simulations. Here, we examine, using the theory underlying the method implemented in the IMA2 program, the possible bases for their observations. We demonstrate that gene migration times are not fully identifiable using the general coalescent for genealogies in an IM model, as implemented in IMA2 and similar programs. In many respects, the findings are general to methods that rely upon calculating the probabilities of genealogies under the coalescent and so are of broader interest than any particular program. We note that the method implemented in IMA2 is the same as that in the IMA program (Hey & Nielsen 2007), and hereafter, we refer simply to IMA.

Correspondence: Vitor C. Sousa, Fax: +1 732 445 6920; E-mail: sousa@biology.rutgers.edu

Principles of IMA

The function of IMA is to obtain the posterior density, $h(\Theta \mid X)$, for the parameters Θ of an IM model given data X from one or more loci from two populations (or more than two populations in the case of IMA2) (for details see Hey & Nielsen 2007; Hey 2010). The parameters Θ include the effective population sizes, migration rates and times of population separation. Hey & Nielsen (2007) showed that the posterior of the parameters $h(\Theta \mid X)$ can be approximated given a sample of genealogies from the posterior density $h(G \mid X)$. In effect, the method collects the information that the data contains about Θ in the form of a sample of genealogies and then uses these genealogies to estimate the posterior density for Θ, i.e. $p(\Theta \mid G,X) = p(\Theta \mid G)$ if $G \sim h(G \mid X)$ (Hey & Nielsen 2007; Hey 2010). But because there is additional information in the genealogies, which does not bear directly on Θ, it is also possible to estimate a posterior density for other quantities, such as the time of most recent common ancestor in the genealogy (TMRCA), the number and time of coalescent events in each population, as well as the number and time of migration events between pairs of populations for each locus. Thus, even though the IM model assumes a constant rate of gene flow since population splitting, it seemed that by examining the genealogies sampled from the posterior density, it would also be possible to estimate the posterior density of migration times (Won & Hey 2005). As Strasburg & Rieseberg (2011) discovered by simulation and as we show here using an approach based on the calculation of the probability of a genealogy, this is not the case.

In IMA and related programs, a value of G is an ultrametric binary tree that depicts the topology, branch lengths, migration times and migration directions for a sample of genes at a locus (Beerli & Felsenstein 1999; Nielsen & Wakeley 2001). To address the identifiability of migration times, we partition G into several compo-

nents, including a topology λ, a vector with the coalescent times $\mathbf{t_c} = (t_{c_1},\ldots,t_{c_T})$, a vector with the migration times $\mathbf{t_m} = (t_{m_1},\ldots,t_{m_T})$, where c_T and m_T are the total number of coalescent and migration events, respectively, and a matrix $\mathbf{n}$, where n_{ji} is the number of lineages in population j at the ith interval between any two events. For simplicity, we refer to the topology and coalescent times as $\Lambda = (\lambda,\mathbf{t_c})$.

The probability of a genealogy, $\pi(G\,|\,\Theta) = \pi(\mathbf{t_m},\mathbf{n},\Lambda\,|\,\Theta)$, is obtained based on coalescent theory assuming a demographic model with parameters Θ. It is noteworthy that $\pi(G\,|\,\Theta)$ does not depend directly on much of the information in a genealogy but rather on a few summaries. In models that include migration, these summaries are counts and sums of rates for coalescent and migration events, including the following: (i) the number of coalescent events in each population $\mathbf{c_c} = (c_{c_1},\ldots,c_{c_p})$; (ii) the number of migration events between each pair of populations $\mathbf{c_m} = (c_{m_{12}},\ldots,c_{m_{p(p-1)}})$; (iii) the sum of coalescent rates for each population $\mathbf{f_c} = (f_{c_1},\ldots,f_{c_p})$; and (iv) the sum of rates for migration events for each pair of populations $\mathbf{f_m} = (f_{m_{12}},\ldots,f_{m_{p(p-1)}})$, where p refers to the number of populations. In more detail, the sums of rates of coalescent for population j and rates of migration between population j and l are defined as functions of the time intervals and number of lineages in each population during each interval:

$$\begin{aligned} f_{c_j}(\mathbf{t},\mathbf{n}) &= \sum_{i=1}^{c_T+m_T}\left[\Delta t_i \binom{\mathrm{n}_{j_i}}{2}\right], \\ f_{m_{jl}}(\mathbf{t},\mathbf{n}) &= \sum_{i=1}^{c_T+m_T}[\Delta t_i \mathrm{n}_{j_i}], \end{aligned} \tag{1}$$

where $\Delta t_i = t_{i+1} - t_i$ is the time interval between any two events, either a coalescent or migration, and $\mathbf{t} = (\mathbf{t_m},\mathbf{t_c})$ is a vector with the sorted coalescent and migration times. For simplicity, these summaries will be referred to as $\mathbf{s} = (\mathbf{c_c},\mathbf{c_m},\mathbf{f_c},\mathbf{f_m})$. For instance, for an IM model, during a time period with p populations, given the scaled effective sizes θ and migration rates m, this probability is

$$\pi(\mathbf{t_m},\mathbf{n},\Lambda|\Theta) = \prod_{j=1}^{p}\left(\frac{2}{\theta_j}\right)^{\mathbf{c}_{c_j}} \mathrm{e}^{\left(-\frac{2}{\theta_j}\mathbf{f}_{c_j}\right)} \prod_{l\neq j} m_{j\to l}^{\mathbf{c}_{m_{jl}}} \mathrm{e}^{(-m_{j\to l}\mathbf{f}_{m_{jl}})}, \tag{2}$$

(Kuhner *et al*. 1998; Beerli & Felsenstein 1999; Hey & Nielsen 2007), where $\theta_j = 4N_{e_j}\mu$, $m_{j\to l} = M_{j\to l}/\mu$, $4N_{e_j}$ is the effective size of population j, μ the mutation rate, and $M_{j\to l}$ is the migration rate between population j and l. Note that the terms following the first and second products are associated with coalescent and migration events, respectively. From eqn 2, we can see that the probability of the genealogy (represented by its components $\mathbf{t_m}$, $\mathbf{n}$, and Λ) depends on the values of the summaries $\mathbf{s} = (\mathbf{c_c},\mathbf{c_m},\mathbf{f_c},\mathbf{f_m})$. All genealogies whose $\mathbf{t_m}$, $\mathbf{n}$ and Λ correspond to the same set of summaries $\mathbf{s}$ have the same prior probability. This is a general result, as eqn 2 is the basis of most inference methods based on genealogies (e.g. Beerli & Felsenstein 1999), including methods where the prior probability of the genealogy is calculated by integrating over the prior distribution of the parameters Θ (Hey & Nielsen 2007; Hey 2010).

As a consequence, for the final step of the estimation of the posterior probability $h(\Theta\,|\,X)$, we can use a sample of values of $\mathbf{s}$ from the posterior of genealogies. The result is a function that is itself a mean of functions, one for each sampled value of $\mathbf{s}$,

$$h(\Theta|X) \approx \frac{1}{k}\sum_{i=1}^{k}\frac{f(\mathbf{s}_i|\Theta)\pi(\Theta)}{\pi(\mathbf{s}_i)}, \tag{3}$$

for a sample of k values of $\mathbf{s}\sim h(\mathbf{s},\Lambda\,|\,X)$, where $\pi(\Theta)$ is the prior of the parameters (Hey & Nielsen 2007; Hey 2010). As $f(\mathbf{s}_i\,|\,\Theta) = f(G_i\,|\,\Theta)/p(G_i\,|\,\mathbf{s}_i)$ and $\pi(\mathbf{s}_i) = \pi(G_i)/p(G_i\,|\,\mathbf{s}_i)$ (similar to eqn A.2), the above expression is an alternative representation for the posterior $h(\Theta\,|\,X)$, which is typically expressed as a function of genealogies (see eqns 11 and 19 in Hey & Nielsen (2007)). In the case of an IM model with two sampled populations and one ancestral population, $\mathbf{s}$ includes just 10 quantities regardless of the sample sizes, and yet, it is sufficient for calculating the probability of a genealogy under the IM model. For multiple independent loci each with a genealogy, $\mathbf{s}$ still includes just 10 quantities, each the sum of the corresponding quantities calculated for the individual loci (Hey & Nielsen 2007; Hey 2010).

Posterior probability of migration times

The posterior probability for the genealogy includes that for the migration times, $\mathbf{t_m}$,

$$h(G|X) = h(\mathbf{t_m},\mathbf{n},\Lambda|X) = f(X|\mathbf{t_m},\mathbf{n},\Lambda)\pi(\mathbf{t_m},\mathbf{n},\Lambda)/f(X), \tag{4}$$

where $f(X\,|\,\mathbf{t_m},\mathbf{n},\Lambda)$ is the likelihood, $\pi(\mathbf{t_m},\mathbf{n},\Lambda)$ is the prior of the genealogy, and $f(X)$ is the marginal likelihood. It is noteworthy that the likelihood depends only on the topology and coalescent times of the genealogy and does not depend on the number and times of migration events (Felsenstein 1988), i.e.

$$f(X|\mathbf{t_m},\mathbf{n},\Lambda) = f(X|\Lambda). \tag{5}$$

This raises the question of whether data can in fact contain any information about the migration times, when considered under an IM model. This can be answered by looking further at the posterior distribu-

tion. Combining eqn 5 in eqn 4 and noting that $h(\Lambda \mid X) = f(X \mid \Lambda)\pi(\Lambda)/f(X)$, the posterior becomes

$$h(\mathbf{t_m}, \mathbf{n}, \Lambda|X) = h(\Lambda|X)\pi(\mathbf{t_m}, \mathbf{n}|\Lambda). \quad (6)$$

This shows that the posterior distribution for the times of migration depends on the posterior for the topology and coalescent times $h(\Lambda \mid X)$ and on the conditional prior $\pi(\mathbf{t_m},\mathbf{n} \mid \Lambda)$ (eqn 6). It can be seen that the most likely migration times are supported by the data indirectly through the posterior of the topology and coalescent times, i.e., the most likely Λ induce a change in the prior of migration timing $\pi(\mathbf{t_m},\mathbf{n} \mid \Lambda)$. This demonstrates that data provide at least some information about the migration timing (eqn 6). However, as we describe later, the data inform us about the most likely values for summaries of the time intervals **s**, rather than about the elements of the migration time vector $\mathbf{t_m}$.

Nonidentifiability of genealogies

Consider two genealogies $G = (\mathbf{t_m},\mathbf{n},\Lambda)$ and $G^* = (\mathbf{t_m}^*,\mathbf{n},\Lambda)$ that share the same coalescent times and topologies, Λ, and the same number of lineages **n** (implying the same number of migrations), but have different migration times, $\mathbf{t_m}$ and $\mathbf{t_m}^*$, respectively. Because the likelihood depends only on Λ (eqn 5) and does not depend on $\mathbf{t_m}$, the posterior probabilities are equal if the two genealogies have the same prior probabilities, $\pi(\mathbf{t_m}, \mathbf{n}, \Lambda) = \pi(\mathbf{t_m}^*, \mathbf{n}, \Lambda)$,

$$h(G|X) = \frac{(X|\Lambda)\pi(\mathbf{t_m}, \mathbf{n}, \Lambda)}{f(X)} = \frac{f(X|\Lambda)\pi(\mathbf{t_m}^*, \mathbf{n}, \Lambda)}{f(X)} = h(G^*|X). \quad (7)$$

As seen in eqn 2, this holds true for genealogies with the same set of summaries **s**. Therefore, it is possible to show that **s** is sufficient for $(\mathbf{t_m},\mathbf{n})$, in the sense that the posterior of the genealogy depends on **s**, irrespective of the particular values of $(\mathbf{t_m},\mathbf{n})$ (see Appendix I). In other words, the posterior of the migration timing (eqn 6) is fully characterized by the posterior $h(\mathbf{s},\Lambda \mid X)$. This means that information provided by the data about the most likely times of migration is captured through the posterior of the summaries **s**. This makes sense because two of the set of summaries ($\mathbf{f_c}$ and $\mathbf{f_m}$) are functions of the time intervals (eqn 1). However, the fact that these summaries are sums of counts and rates of events across loci introduces an identifiability problem. The reason is that we can estimate the most likely values for the sums given the data, $h(\mathbf{s},\Lambda \mid X)$, but we cannot expect to estimate each term of the sum. In particular, there are multiple combinations of $(\mathbf{t_m},\mathbf{n})$ for a given value of **s**. Therefore, we can have two or more genealogies with the same posterior probability but with different migration timing distributions. In these cases, genealogies are said to be nonidentifiable as it is impossible to distinguish them based on their posterior.

Figure 1 shows an example of this nonidentifiability using two genealogies with different migration timings. In the left panel, both migrations happen recently, whereas in the right panel, both migrations happen just after the population split. Despite having different migration times, both genealogies have the same values for the summaries $\mathbf{s} = (\mathbf{c_c},\mathbf{c_m},\mathbf{f_c},\mathbf{f_m})$ and for the coalescent time $\mathbf{t_c}$, and hence have the same posterior probabilities. As seen in the Fig. 1, all genealogies with the same time interval Δt and $\mathbf{t_c}$ have the same posterior, despite having different migration timing $\mathbf{t_m}$.

When there are multiple loci, the nonidentifiability issue is compounded because the posterior probability of all the genealogies depends on summaries that are the sums of **s** for each of the individual loci. Figure 2 shows an example for two loci. As can be seen, genealogies have migrations in different periods of time, which are consistent in both loci. In Fig. 2a, the two loci suggest older migration, whereas in Fig. 2b, the two loci have recent migration events. These two different cases could be interpreted as favoring alternative models of divergence, if it were possible to distinguish them. But because **s** is a sum over loci, given that in this example $(\Delta t_1+\Delta t_2) = (\Delta t_1{}^*+\Delta t_2{}^*)$ and the coalescent times are the same, the two groups of genealogies will have the same value of **s**. Hence, these two groups of genealogies have the same posterior, despite the very different times of migration.

Relation between genealogy summaries and migration times

Given that some information about migration time is contained in the data (eqn 6), we wondered if some general feature of the migration times are contained in **s**, particularly the summary $\mathbf{f_m}$ that is the sum of migration rates over time intervals (eqn 1). Data sets were simulated and the joint distribution of $\mathbf{f_m}$ and overall measures of migration, including the mean, minimum and maximum migration time, were recorded. Simulations were carried out under an IM model, which assumes a constant migration rate, with two sampled populations that diverged from one ancestral population, using the coalescent-based simulator implemented in SIMDIV (Wang and Hey 2010). Data sets were generated with a fixed set of parameter values ($\theta_1 = \theta_2 = \theta_A = 5.0$, $m_{1\to 2} = m_{2\to 1} = 0.5$ and $t_{\text{split}} = 2.0$), varying the sample sizes in each population $n = (2,10,100)$. If genealogies contain information about these overall

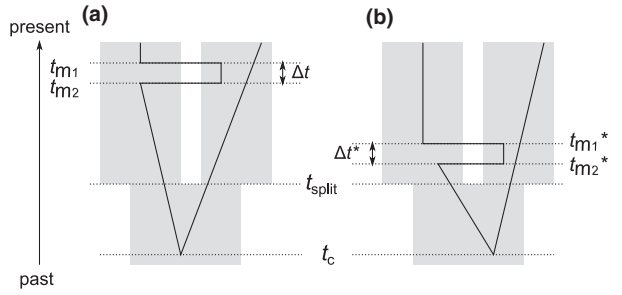

Fig. 1 Example of nonidentifiability of migration timing for single-locus genealogies. Genealogies under an IM model for two gene copies from two sampling populations and an ancestral population. Both genealogies share the same time of population split t_{split}, topology and coalescent time Λ, number of migrations $\mathbf{c_m}$ and number of coalescent events $\mathbf{c_c}$, but have different migration times, $\mathbf{t_m}$ and $\mathbf{t_m}^*$, respectively. Thus, the two genealogies have the same values for $\mathbf{c_c} = (0,0,1)$, $\mathbf{c_m} = (1,1)$. If the time interval between the two migration events, Δt, is the same in both genealogies, they will also have the same values for $(\mathbf{f_c},\mathbf{f_m})$. For instance, if $\Delta t = \Delta t^* = 2$, $t_{split} = 10$ and $t_c = 15$, then $\mathbf{f_c} = (0,2,5)$ and $\mathbf{f_m} = (8,12)$ for both genealogies. Both genealogies have the same summaries $\mathbf{s}$ and hence have the same posterior probability (eqn 7).

measures of migration time, then we would expect to see a correlation with $\mathbf{f_m}$. However, as shown in Fig. 3, this was not observed. Regardless of sample size, $\mathbf{f_m}$ shows only a quite modest association with the mean, minimum or maximum of $\mathbf{t_m}$. The Spearman's rank correlation coefficients were low, ranging from 0.09 to 0.12 for the mean, from 0.07 to 0.10 for the maximum, and from 0 to 0.05 for the minimum. Similar results were obtained for $\mathbf{f_c}$ (not shown). These results suggest that we cannot expect to estimate these features of $\mathbf{t_m}$.

Discussion

Strasburg & Rieseberg (2011) demonstrated with simulations an identifiability problem for migration timing. Here, we explain the underlying basis of their findings in terms of the calculation for the probability of genealogies. When using the coalescent to calculate the probability of genealogies under a model with migration, such as the IM model, the probability of a genealogy depends only on a modest set of summaries $\mathbf{s} = (\mathbf{c_c},\mathbf{c_m},\mathbf{f_c},\mathbf{f_m})$ (Hey & Nielsen 2007), which means that genealogies that differ in their times of migration can have the same values for $\mathbf{s}$. This implies that genealogies with different migration timings can have the same posterior probability and that the migration timings are statistically nonidentifiable. Investigators cannot expect to be able to estimate migration times for the purpose of discerning models of population or species divergence where gene flow varies through time.

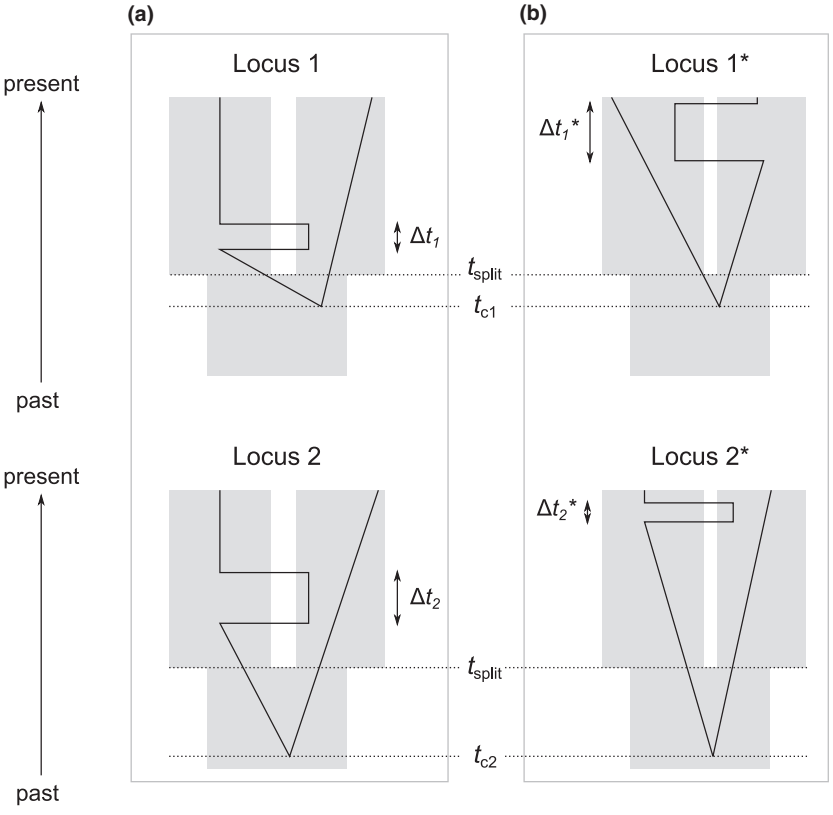

Fig. 2 Nonidentifiability of migration timing for multiple loci. Example of two sets of genealogies for two loci with different times of migration, but with the same time of split t_{split} and coalescent times t_{c_1} and t_{c_2}, for locus 1 and locus 2, respectively. The posteriors are $h(G_1,G_2 \mid X)$ and $h(G_1^*,G_2^* \mid X)$ for (a) and (b), respectively. Given that the summaries are summed over loci, the two posterior distributions are the same if $(\Delta t_1+\Delta t_2) = (\Delta t_1^*+\Delta t_2^*)$. For instance, with $(\Delta t_1+\Delta t_2) = (\Delta t_1^*+\Delta t_2^*) = 8$, $t_{split} = 10$, $t_{c_1} = 12$ and $t_{c_2} = 15$, the summaries are $\mathbf{c_c} = (0,0,2)$, $\mathbf{c_m} = (2,2)$, $\mathbf{f_c} = (0,8,7)$ and $\mathbf{f_m} = (12,28)$ for both sets of genealogies. Provided that the summaries and times of coalescent are the same, the two posterior distributions are identical $h(G_1,G_2 \mid X) = h(G_1^*,G_2^* \mid X)$. Note that the summaries $\mathbf{f_c}$ and $\mathbf{f_m}$ depend on the time intervals, rather than in the actual times of migration $\mathbf{t_m}$. See legend of Fig. 1 and text for details.

This is a general result applicable to genealogies under neutral demographic models that include migration and that depend on the coalescent theory. We thus expect that migration timing estimates obtained with programs such as MDIV (Nielsen & Wakeley 2001), IMA (Hey & Nielsen 2004, 2007), LAMARC (Kuhner *et al.* 1998; Kuhner 2006) and MIGRATE (Beerli & Felsenstein 1999) will suffer from this limitation. It is noteworthy that the nonidentifiability of migration timing does not introduce any bias in the estimates of the demographic parameters, such as the effective sizes and migration rates, because the summaries capture all the genealogical information needed to estimate the posterior of the parameters (eqn 3) (Hey & Nielsen 2007; Hey 2010).

Previous studies have reported a wide range of shapes for the posterior distribution of migration timings, including cases suggesting recent migrations, old migrations and/or complex multimodal distributions (e.g. Niemiller *et al.* 2008; Strasburg *et al.* 2008; Nadachowska & Babik 2009; Carneiro *et al.* 2010). The presence of a peak and of variation in the number and location of peaks in the posterior distribution lends the appearance that these distributions are informative. However, this is misleading as the estimated posterior densities for migration times are mostly a function of (i) the prior distribution of migration times and (ii) the nonidentifiability problem. Unlike the prior distributions for the migration rates that are usually uniform and specified by the investigator, the prior distributions for the migration times are induced by the model assumptions. In a model with constant gene flow, the prior distribution for the migration times is not expected to be uniform, but rather a decreasing function with a peak close to zero. The reason is that the number

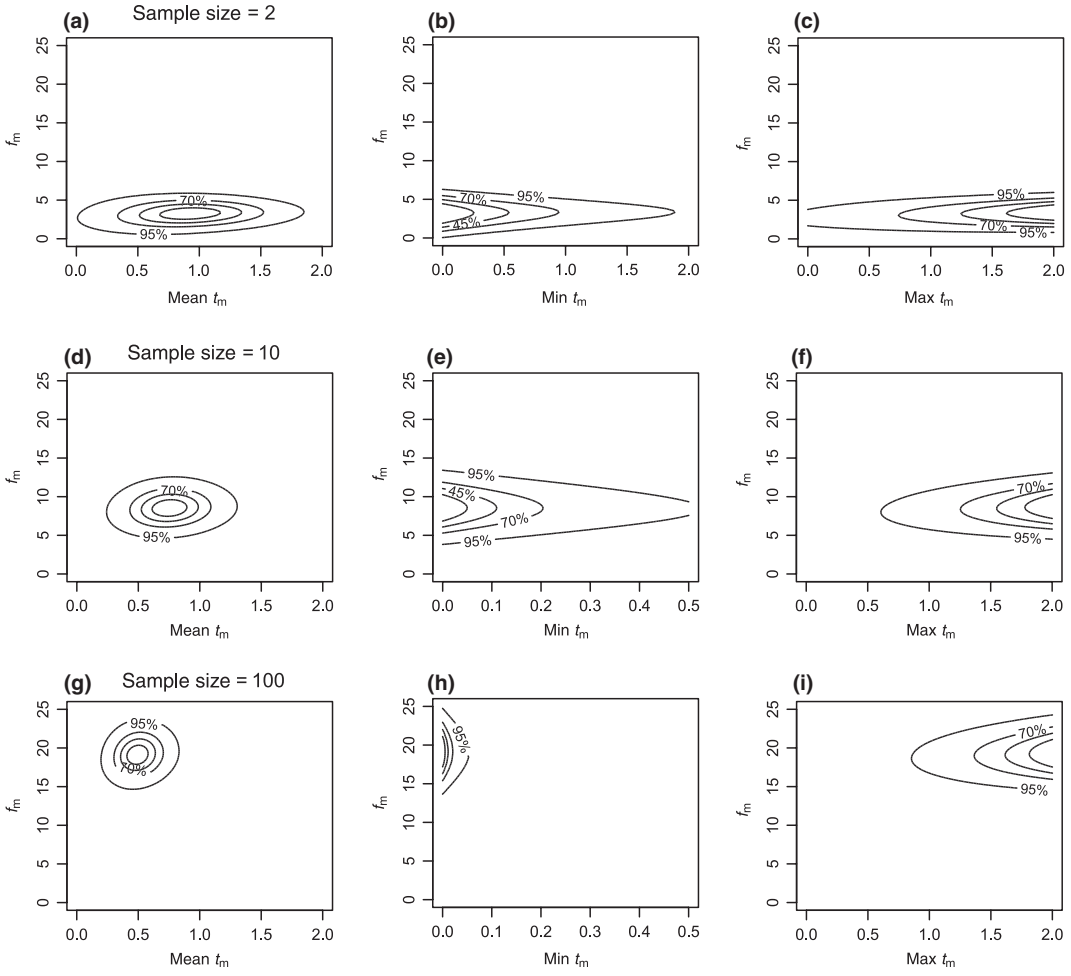

Fig. 3 Contour plots of the joint distributions of the mean, minimum and maximum of $\mathbf{t_m}$ and the summary of the genealogy $f_{m_{12}}$. These were obtained with 50 000 simulations under a two population IM model with parameters $\theta_1 = \theta_2 = \theta_A = 5.0$, $m_{1 \to 2} = m_{2 \to 1} = 0.5$ and $t_{split} = 2.0$. (a–c) Results obtained with a sample size of two gene copies in each population; (d–f) Results obtained with a sample size of 10 gene copies in each population; (g–i) Results obtained with a sample size of 100 gene copies in each population. Note that these plots correspond to empirical prior distributions obtained with simulations and not to posterior distribution estimated using IMA.

of migration events is proportional to the number of lineages in each population at any instant, and given that the number of lineages decreases going backwards in time owing to coalescent events, most migrations are expected to occur recently. This may explain some of the results found suggesting recent migration. In addition, the effects of the nonidentifiability on the posteriors arise because of the fact that the summaries **s** are sufficient (eqn A.1) and sums of functions of the migration and coalescent times (eqn 1). Given a particular data set, the most likely values for the summaries **s** impose strong correlations on the migration times $\mathbf{t_m}$. The shape of the posteriors is thus a function of the correlations between the migration times, which depend on the information contained in the data about the values of the summaries. This is influenced by the properties of each particular dataset, such as the sample sizes, sequence lengths, number of loci, as well as the priors specified for the demographic parameters. As a consequence, the posteriors can have complex shapes, including distributions with multiple peaks. In any case, the fact that the times of migration are nonidentifiable implies that the posterior distributions do not have the desirable property of identifying the correct times of migration. Thus, irrespective of its shape, these are not useful to estimate the times of migration.

The initial motivation for looking at the posterior of migration timing was to infer variation in gene flow through time (e.g. Won & Hey 2005). As noted by Strasburg & Rieseberg (2011), cases in which the migration rates vary through time violate the assumptions of the basic IM model. We can envision at least two possible approaches to modelling variable migration rates explicitly. One is to assume that migration rates vary through time following some deterministic function, e.g., exponential change, the parameters of which are estimated from the data along with other parameters. Another possibility is to include in the model more migration parameters, each associated with a distinct time period (e.g. as used in simulations by Becquet & Przeworski 2009). In the simplest case of an IM model with two sampled populations, there would be two migration periods, each with its own migration rates, as well as an additional parameter for the time at which migration rate changed. However, this approach increases significantly the number of parameters of the model, and it is possible that a large amount of additional data would be required for estimation.

Acknowledgements

We thank three anonymous reviewers for their comments. This work was supported by the National Science Foundation (NSF) grant DEB-0949561 and by National Institutes of Health (NIH) grant GM078204 to J.H.

References

Becquet C, Przeworski M (2009) Learning about modes of speciation by computational approaches. *Evolution*, **63**, 2547–2562.

Beerli P, Felsenstein J (1999) Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, **152**, 763–773.

Carneiro M, Blanco-Aguiar J, Villafuerte R, Ferrand N, Nachman M (2010) Speciation in the European rabbit (*Oryctolagus cuniculus*): islands of differentiation on the X chromosome and autosomes. *Evolution*, **64**, 3443–3460.

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

Hey J (2010) Isolation with migration models for more than two populations. *Molecular Biology and Evolution*, **27**, 905–920.

Hey J, 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, Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences USA*, **104**, 2785–2790.

Kuhner M (2006) LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics*, **22**, 768–770.

Kuhner M, Yamato J, Felsenstein J (1998) Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, **149**, 429–434.

Lehmann E, Casella G (1998) *Theory of Point Estimation*. Springer Verlag, Berlin.

Nadachowska K, Babik W (2009) Divergence in the face of gene flow: the case of two newts (Amphibia: Salamandridae). *Molecular Biology and Evolution*, **26**, 829–841.

Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.

Niemiller M, Fitzpatrick B, Miller B (2008) Recent divergence with gene flow in Tennessee cave salamanders (Plethodontidae: *Gyrinophilus*) inferred from gene genealogies. *Molecular Ecology*, **17**, 2258–2275.

Pinho C, Hey J (2010) Divergence with gene flow: Models and data. *Annual Review of Ecology, Evolution, and Systematics*, **41**, 215–230.

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

Strasburg J, Rieseberg L, Kohn J (2008) Molecular demographic history of the annual sunflowers *Helianthus annuus* and *H. petiolaris*-Large effective population sizes and rates of long-term gene flow. *Evolution*, **62**, 1936–1950.

Wang Y, Hey J (2010) Estimating divergence parameters with small samples from a large number of loci. *Genetics*, **184**, 363–379.

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

J.H. conducts empirical and theoretical genetic research on diverse problems in speciation and evolutionary genetics. A.G. and V.C.S. are postdoctoral fellows at the Hey lab working on the population genetics of diverging populations and development of statistical methods.

Appendix I

Here, we demonstrate that the summaries of the genealogy $\mathbf{s} = (\mathbf{c_c}, \mathbf{c_m}, \mathbf{f_c}, \mathbf{f_m})$ are sufficient for the migration timing $\mathbf{t_m}$ and number of lineages $\mathbf{n}$. This is analogous to demonstrating that a given statistic is sufficient for the parameters of a model. Note that by definition, a statistic is a function of the data, whereas we are dealing with functions of genealogies. This can be shown applying the factorisation theorem (Lehmann & Casella 1998) to the posterior

$$h(\mathbf{t_m}, \mathbf{n}, \Lambda | X) = p(\mathbf{t_m}, \mathbf{n} | \Lambda, \mathbf{s}) h(\mathbf{s}, \Lambda | X), \tag{A.1}$$

where $p(\mathbf{t_m}, \mathbf{n} \mid \Lambda, \mathbf{s})$ is the probability of $(\mathbf{t_m}, \mathbf{n})$ given the values of $\mathbf{s}$, and $h(\mathbf{s}, \Lambda \mid X)$ is the posterior of $\mathbf{s}$. Noting that $h(\mathbf{t_m}, \mathbf{n}, \Lambda \mid X) = h(\Lambda \mid X)\pi(\mathbf{t_m}, \mathbf{n} \mid \Lambda)$ (eqn 6) and that $h(\mathbf{s}, \Lambda \mid X) = h(\Lambda \mid X)\pi(\mathbf{s} \mid \Lambda)$, the above-mentioned equation becomes

$$\pi(\mathbf{t_m}, \mathbf{n}, \Lambda) = p(\mathbf{t_m}, \mathbf{n} | \mathbf{s}, \Lambda) \pi(\mathbf{s}, \Lambda). \tag{A.2}$$

Thus, showing that the prior $\pi(\mathbf{t_m}, \mathbf{n}, \Lambda)$ can be factorized into the two functions $p(\mathbf{t_m}, \mathbf{n} \mid \mathbf{s}, \Lambda)$ and $\pi(\mathbf{s}, \Lambda)$, implies that $\mathbf{s}$ is sufficient for the posterior $h(\mathbf{t_m}, \mathbf{n} \mid X)$. The function $p(\mathbf{t_m}, \mathbf{n} \mid \mathbf{s}, \Lambda)$ reflects the probability of obtaining a given configuration for $(\mathbf{t_m}, \mathbf{n})$ conditional on the values of the summaries $\mathbf{s}$. Note that it does not depend on the data X as required for $\mathbf{s}$ to be considered sufficient. Given that all genealogies that have the same corresponding values for the summaries are equally likely (eqn 2), the probability $p(\mathbf{t_m}, \mathbf{n} \mid \mathbf{s}, \Lambda)$ will be proportional to the number of genealogies sharing the same values for $\mathbf{s}$.

The prior $\pi(\mathbf{s}, \Lambda)$ is obtained by integrating over the prior probability of genealogies whose $(\mathbf{t_m}, \mathbf{n}, \Lambda)$ correspond to a given set of summaries $\widetilde{\mathbf{s}}$,

$$\pi(\mathbf{s} = \widetilde{\mathbf{s}}, \Lambda) = \int \pi(\mathbf{t}_m, \mathbf{n}, \Lambda) \mathbb{1}_{\{\mathbf{s}(\mathbf{t_m}, \mathbf{n}) = \widetilde{\mathbf{s}}\}} \mathrm{d}\mathbf{t_m} \mathrm{d}\mathbf{n}, \tag{A.3}$$

where $\mathbb{1}_{\{c\}}$ is an indicator variable that takes the value 1 if the condition c holds true and zero otherwise. The same reasoning applies to the posterior $h(\mathbf{s}, \Lambda \mid X)$. Again, note that $h(\mathbf{s}, \Lambda \mid X)$ does not depend on $(\mathbf{t_m}, \mathbf{n})$, as required for $\mathbf{s}$ to be considered sufficient. Given that $\mathbf{s}$ is sufficient and a sum of counts and rates across period of the genealogy and across loci, the elements of the sum $(\mathbf{t_m}, \mathbf{n})$ are nonidentifiable.