Inference of gene flow in the process of speciation: efficient maximum-likelihood implementation of a generalised isolation-with-migration model

Rui J. Costa* and Hilde M. Wilkinson-Herbots†

Department of Statistical Science, University College London,

Gower Street, London WC1E 6BT, UK

Abstract

The 'isolation with migration' (IM) model has been extensively used in the literature to detect gene flow during the process of speciation. In this model, an ancestral population split into two or more descendant populations which subsequently exchanged migrants at a constant rate until the present. Of course, the assumption of constant gene flow until the present is often over-simplistic in the context of speciation. In this paper, we consider a 'generalised IM' (GIM) model: a two-population IM model in which migration rates and population sizes are allowed to change at some point in the past. By developing a maximum-likelihood implementation of this model, we enable inference on both historical and contemporary rates of gene flow between two closely related populations or species. The GIM model encompasses both the standard two-population IM model and the 'isolation with initial migration' (IIM) model as special cases, as well as a model of secondary contact. We examine for simulated data how our method can be used, by means of likelihood ratio tests or AIC scores, to distinguish between the following scenarios of population divergence: (a) divergence in complete isolation; (b) divergence with a period of gene flow followed by isolation; (c) divergence with a period of isolation followed by secondary contact; (d) divergence with ongoing gene flow. Our method is based on the coalescent and is suitable for data sets consisting of the number of nucleotide differences between one pair of DNA sequences at each of a large number of independent loci. As our method relies on an explicit expression for the likelihood, it is computationally very fast.

Keywords: speciation, coalescent, maximum-likelihood, gene flow, isolation, secondary contact.

*Present address: European Bioinformatics Institute (EMBL-EBI), Hinxton, UK.

†Corresponding author.

E-mail addresses: ruibarrigana@ebi.ac.uk (R. J. Costa), h.herbots@ucl.ac.uk (H. M. Wilkinson-Herbots).

1 Introduction

Molecular genetic data have been used extensively to learn about the evolutionary processes that gave rise to the observed genetic variation. One important example is the use of genetic data to try to infer whether or not gene flow occurred between closely related species during or after speciation. Such studies have often used computer programs such as MDIV (Nielsen and Wakeley, 2001), IM (Hey and Nielsen, 2004; Hey, 2005), IMa (Hey and Nielsen, 2007), MIMAR (Becquet and Przeworski, 2007) or IMa2 (Hey, 2010), based on the 'isolation with migration' (IM) model, which assumes that a panmictic ancestral population instantaneously splits into two or more descendant populations, which subsequently exchange migrants at a constant rate until the present. A meta-analysis of research papers that have used the IM model in the context of speciation can be found in Pinho and Hey (2010).

While the above methods were aimed at data from a large number of individuals at a relatively small number of loci and are computationally intensive, advances in DNA sequencing technology and the advent of whole-genome sequencing have led to an increased interest in methods which are able to detect gene flow using data at large numbers of loci but from pairs or small numbers of sequences. We will focus here on maximum-likelihood (ML) methods for such data, which typically assume that there is no recombination within loci and free recombination between loci. This type of data set has two advantages. Firstly, data from even a very large number of individuals from the same population at the same locus tend to contain only little information about very old events, since the individuals' ancestral lineages will typically have coalesced to a very small number of ancestral lineages by the time the event of interest is reached; in such contexts, a data set consisting of a small number of DNA sequences at each of a large number of independent loci is likely to be more informative (Maddison and Knowles, 2006; Wang and Hey, 2010; Lohse et al., 2010, 2011). Secondly, considering a small number of sequences at each of many independent loci is mathematically much easier and computationally much faster than working with large numbers of sequences at the same locus. In particular, explicit analytical expressions for the likelihood have been obtained for pairs or small numbers of sequences for a number of demographic models (for example, Takahata et al., 1995; Wilkinson-Herbots, 2008; Hobolth et al., 2011; Lohse et al., 2011; Wilkinson-Herbots, 2012; Zhu and Yang, 2012; Andersen et al.,

2014; Lohse and Frantz, 2014; Lohse et al., 2016; Costa and Wilkinson-Herbots, 2017; Dalquen et al., 2017), which can hugely speed up the computation and maximization of the likelihood. Methods of maximum-likelihood estimation of the parameters of the IM model, suitable for small numbers of sequences at each of a large number of independent loci, were developed by Wilkinson-Herbots (2008), Wang and Hey (2010), Hobolth et al. (2011), Lohse et al. (2011), Zhu and Yang (2012), Andersen et al. (2014), and Dalquen et al. (2017). Because the IM model – and in particular its assumption of migration continuing at a constant rate until the present – is clearly unrealistic in the context of speciation, these methods have also been modified or extended to incorporate some forms of temporal changes in migration rates. Innan and Watanabe (2006) implemented a model in which gene flow between two diverging species decreases linearly with time and eventually ceases. While their model is more sophisticated than some of the later models discussed here, their calculation of the likelihood of the number of nucleotide differences between pairs of sequences relies on the numerical computation of the coalescence time density using recursion equations on a series of time points, and hence is not as fast as methods that use explicit analytical expressions for the likelihood. Lohse and Frantz (2014) implemented a computationally efficient ML method for a model of introgression between two species where an instantaneous admixture event occurred at a single point in time; see also Hearn et al. (2014). Lohse et al. (2011, 2016) also developed a more general Laplace transform method to calculate blockwise likelihoods for a range of demographic scenarios. In our previous work (Wilkinson-Herbots, 2012; Wilkinson-Herbots, 2015; Costa and Wilkinson-Herbots, 2017) we developed a fast ML method for an 'isolation with initial migration' (IIM) model in which two diverging populations experience gene flow at a constant rate for a period of time and subsequently become completely isolated.

In recent years there has been intense interest in inferring not only whether or not gene flow occurred during the process of speciation, but also in distinguishing divergence in the face of gene flow from secondary contact and, more generally, distinguishing decreasing from increasing gene flow (see, for example, Roux et al., 2016, and the review by Sousa and Hey, 2013). The present paper contributes to this aim. We extend our earlier work on the IM and IIM models to the 'generalised isolation with migration' (GIM) model depicted in Figure 1d: this is a two-population IM model but which allows for an instantaneous change of the migration rates and population sizes at some

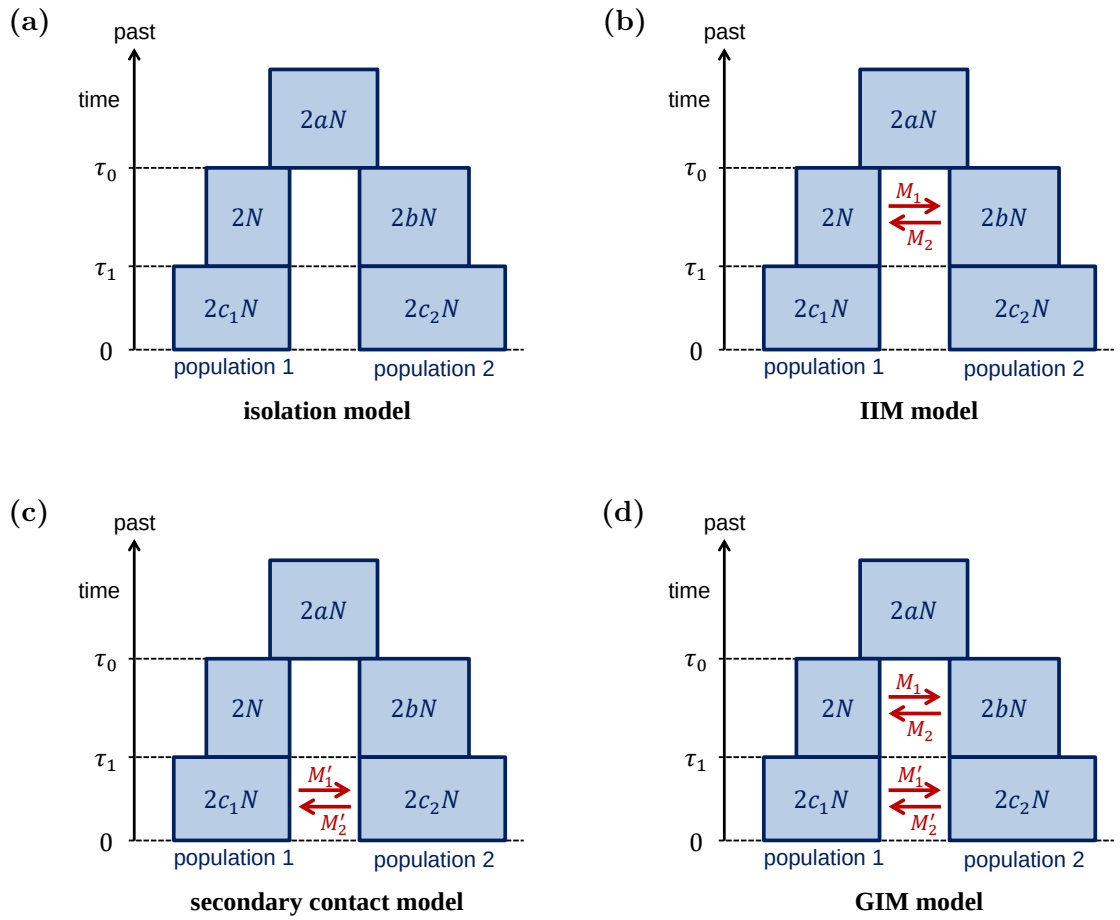

Figure 1: The models of population divergence considered in this paper. Figure (d) depicts the generalised isolation-with-migration (GIM) model which is the main focus of this paper. Figures (a) to (c) represent simpler models nested in the GIM model: (a) an isolation model with a potential change of the descendant population sizes; (b) the isolation-with-initial migration (IIM) model; and (c) a model of secondary contact. All four models assume that an ancestral population of size $2aN$ homologous DNA sequences split into two descendant populations of sizes $2N$ and $2bN$ sequences, time τ_0 ago, which may have undergone a subsequent size change at time τ_1 ago, resulting in populations of sizes $2c_1N$ and $2c_2N$ sequences, respectively. Depending on the model, gene flow may have occurred between times τ_0 and τ_1 ago and/or between time τ_1 ago and the present; M_i and M_i' $(i = 1, 2)$ denote the 'scaled' migration rates backward in time during the time periods indicated in the diagrams.

point in the past. A useful feature of this model is that it encompasses both the IIM model (Figure 1b) and a model of secondary contact (Figure 1c) as special cases, as well as a model of isolation with a possible change of the descendant population sizes (Figure 1a), so that our ML method provides an easy way to quickly compare how well the four evolutionary scenarios in Figure 1 fit a particular data set; in addition, the fit of any other models nested in the GIM model, such as versions with unidirectional or symmetric gene flow, and the original IM and isolation models, can also be compared. Thus it is possible to distinguish between historical and contemporary gene flow, while also distinguishing between the effects of gene flow and those of population size changes.

While the IIM model may be a reasonable (albeit much simplified) description of two gradually diverging species, the model of secondary contact in Figure 1c represents the case of two populations which underwent a period of isolation (for example, due to climatic changes or habitat fragmentation) and subsequently became reconnected by gene flow, whereas two gradually diverging populations which have not yet reached complete reproductive isolation might be described by a GIM model with decreasing gene flow. The models considered in this paper were motivated by our joint work in Janko et al. (2018), where our method (slightly simplified, with symmetric migration rates) was applied to data from pairs of species of *Cobitis* (spined loaches), as part of a broader study examining the interconnection between hybrid asexuality and speciation; however, because the focus of that paper was on various types of biological evidence for the evolutionary processes being studied, it did not include the mathematical results on which our ML method is based, nor did it contain a simulation study examining the performance of our method - these mathematical and computational aspects of our work are presented here.

Our method is applicable to data sets from closely related species or populations, consisting of the numbers of nucleotide differences between one pair of DNA sequences sampled at each of a large number of different loci; for mathematical simplicity and as is common in this context, we will assume that there is no recombination within loci and free recombination between loci. At each locus, the two sampled DNA sequences may be both from descendant population 1, or both from descendant population 2, or one sequence from each of the two descendant populations; a sufficient number of independent loci should be included for each of these three types of pairwise comparisons. Our method was implemented in R (R Core Team, 2019), and our code is available at https://github.com/Costa-and-Wilkinson-Herbots/GIM. Because our ML method is based on an explicit expression for the likelihood, it is computationally very fast. For example, for simulated data from 40,000 loci, computing ML estimates of the 11 parameters of the GIM model typically took about 1 minute of computing time on an ordinary computer, while fitting all four models shown in Figure 1 typically took around 3 minutes of computing time in total.

Models of speciation such as the one implemented in the present paper are defined by a number of assumptions which are at best good approximations of the truth. This holds

for assumptions such as the absence of intra-locus recombination and the lack of linkage between loci, assumptions about the mutation process, and demographic assumptions such as the piecewise constant population sizes and the piecewise constant migration rates. On the issue of how model misspecification can be accounted for when making inferences under a GIM model, we refer the reader to the Discussion and the references therein.

In the previous paragraphs we focused on fast ML methods based on samples of small numbers of DNA sequences from a large number of loci. It should be noted that a number of other, more computationally intensive, methods have been implemented that are able to fit a variety of demographic models, some of which account for recombination as well as gene flow. Notably, Mailund et al. (2012) developed a more complex ML method for an IIM model which accounts for recombination, using a hidden Markov model and the so-called 'Sequential Markov Coalescent' approach. Flouri et al. (2020) developed a full-likelihood Bayesian MCMC implementation of the multi-species coalescent which allows for instantaneous introgression or hybridization events, and which can handle DNA sequence data from a relatively large number of individuals from multiple species at a large number of loci. A different class of methods uses a summary statistic known as the 'site frequency spectrum' of SNP data to fit a range of demographic models, including scenarios with gene flow, by means of a composite-likelihood approach (for example, Gutenkunst et al., 2009; Naduvilezhath et al., 2011; Chen, 2012; Lukić and Hey, 2012; Excoffier et al., 2013; Kern and Hey, 2017). To overcome some of the limitations of such methods (discussed in Terhorst and Song, 2015), Beeravolu et al. (2018) took this approach a step further by developing a simulation-based composite-likelihood method based on the 'blockwise site frequency spectrum' of data consisting of blocks of sequence along the genome, from multiple individuals.

The structure of this paper is as follows. In Section 2 we formulate the GIM model in the context of coalescent theory. We derive an explicit expression for the probability distribution of the number of nucleotide differences between two DNA sequences sampled at random, either both from the same descendant population, or one sequence from each of the two descendant populations; thus we obtain an explicit expression for the likelihood of a data set consisting of the numbers of nucleotide differences between one pair of DNA sequences sampled at each of a large number of independent loci. We also

set out the procedures we will use for model comparison, using either AIC scores or a sequence of likelihood ratio tests. Section 3 contains a simulation study, examining the accuracy of the ML estimates of the parameters of the GIM model obtained with our method, and investigating the results of our model selection procedures, for data sets simulated from a range of different scenarios encompassed by the GIM model. Section 4 contains a discussion of our findings.

2 The generalised isolation-with-migration model

The generalised isolation-with-migration (GIM) model considered in this paper can be described as an isolation-with-migration (IM) model which allows for a change of migration rates and descendant population sizes at some point in the past. It encompasses, as special cases, the standard IM model, the isolation model (with or without a change of descendant population sizes), the isolation-with-initial-migration (IIM) model, and a model of secondary contact.

We assume that, time τ_0 ago ($\tau_0 > 0$), a panmictic ancestral population instantaneously split into two descendant populations which subsequently may have experienced gene flow in one or both directions until the present time. Time τ_1 ago ($0 < \tau_1 < \tau_0$), the population sizes and migration rates may have undergone an instantaneous change. Between times τ_0 and τ_1 ago, and between time τ_1 ago and the present, the migration rates and population sizes are assumed to have been constant. This model is illustrated in Figure 1d. For now, we restrict our attention to DNA sequences at a single locus that is not subject to intralocus recombination. For mathematical convenience and for the sake of consistency with our earlier work (Wilkinson-Herbots, 2008, 2012; Wilkinson-Herbots, 2015; Costa and Wilkinson-Herbots, 2017), the size of descendant population 1 between times τ_0 and τ_1 ago is assumed to be $2N$ sequences, where N is large, and all other population sizes are expressed as fractions or multiples of $2N$. The ancestral population is assumed to have been of constant size $[2aN]$ sequences ($a > 0$) until the split occurred time τ_0 ago, where $[\cdot]$ denotes the integer part function. Between times τ_0 and τ_1 ago, descendant population 2 was of size $[2bN]$ sequences ($b > 0$). From time τ_1 ago until the present, the descendant population sizes were $[2c_1N]$ and $[2c_2N]$ sequences, respectively ($c_1, c_2 > 0$). We further assume that the populations evolve in discrete non-overlapping generations, and that reproduction within each population follows the neutral Wright-

Fisher model (Fisher, 1930; Wright, 1931). Between times τ_0 and τ_1 ago, there may be gene flow between the two descendant populations, at a constant rate in each direction: we assume that in each generation, a fraction $m_i \geq 0$ of descendant population i are immigrants from descendant population j ($i, j \in \{1, 2\}$ with $j \neq i$), i.e. m_i denotes the migration rate per generation from population i to population j backward in time. For the period from time τ_1 ago until the present, the backward migration rates $m_i' \geq 0$ ($i = 1, 2$) are defined analogously. It is assumed that each generation, Wright–Fisher type reproduction within the descendant populations restores them to their stated sizes, i.e., reproduction undoes any decrease or increase in population sizes caused by gene flow. As is standard in coalescent theory, we will measure time in units of $2N$ generations (this also applies to the times τ_0 and τ_1), and we define the 'scaled' migration rates backward in time by $M_i = 4Nm_i$ and $M_i' = 4Nm_i'$, for $i = 1, 2$.

2.1 The coalescent under the GIM model

Tracing back the ancestry of a sample of sequences taken from the present generation (from one or both descendant populations), any two ancestral lineages will coalesce when their most recent common ancestor is reached; lineages can only coalesce when they are in the same population. Working backward in time, the genealogical process of a sample of sequences can be described by a succession of three Markov Chains. Between the present and time τ_1 ago, the genealogy of a sample of sequences is well approximated by the 'structured coalescent' (Takahata, 1988; Notohara, 1990; Herbots, 1997; Kozakai et al., 2016), which is a continuous-time Markov Chain keeping track of the number of distinct ancestral lineages the sample has in each subpopulation, at each time in the past. As time is measured in units of $2N$ generations and N is large, the coalescence rate of any two lineages residing in descendant population i is $1/c_i$, and each lineage moves from descendant population i to descendant population j at rate $M_i'/2$ (for $i, j \in \{1, 2\}$ with $j \neq i$). Between times τ_1 and τ_0 ago, the genealogy of a sample of sequences is again described by the structured coalescent, but now with coalescence rate 1 for any two lineages in descendant population 1, coalescence rate $1/b$ for any two lineages in descendant population 2, and migration rate $M_i/2$ for any lineage in descendant population i. From time τ_0 ago further back into the past, the genealogy of a sample of sequences follows Kingman's coalescent (Kingman, 1982a,b,c), with any

pair of lineages coalescing at rate $1/a$. In what follows, we will refer to the stochastic process described above as the 'coalescent under the GIM model'.

In this paper we will focus on the genealogy of one pair of sequences sampled from the present populations. From the present until time τ_0 ago, this coalescent process has four possible states: state 1, if there are two ancestral lineages in descendant population 1; state 2, if there are two lineages in descendant population 2; state 3, if there is one lineage in each descendant population; and state 4, if coalescence has occurred. Beyond time τ_0 into the past, there are only two possible situations: either there are two distinct ancestral lineages or coalescence has occurred. However, to facilitate the derivation of the coalescence time (the time since the most recent common ancestor of the two sampled sequences), we let the process have four states even beyond time τ_0 into the past: if the coalescent process reaches time τ_0 in state i ($i \in \{1, 2, 3\}$), then the process remains in that state until the two lineages coalesce, at which time the process moves to state 4. The coalescent process thus forms a non-homogeneous continuous-time Markov Chain with state space $\{1, 2, 3, 4\}$, and with piecewise constant transition rates which change at times τ_1 and τ_0. The process starts in state 1, 2 or 3, depending on whether two DNA sequences are sampled both from descendant population 1, both from descendant population 2, or one sequence from each descendant population. The process is absorbed when the two ancestral lineages coalesce at their most recent common ancestor, i.e. when the process reaches state 4. We will derive the distribution of the coalescence time, T_i, of the two sampled sequences, i.e. the distribution of the time until the coalescent process is absorbed into state 4, starting from state i, for $i = 1, 2, 3$.

Formally, for a sample of two sequences, the coalescent under the GIM model is defined by the following three infinitesimal generator matrices. When $0 \leq t \leq \tau_1$,

$$\mathbf{Q_1} = \begin{bmatrix} -\left(\frac{1}{c_1} + M_1'\right) & 0 & M_1' & \frac{1}{c_1} \\ 0 & -\left(\frac{1}{c_2} + M_2'\right) & M_2' & \frac{1}{c_2} \\ \frac{M_2'}{2} & \frac{M_1'}{2} & -\left(\frac{M_1'+M_2'}{2}\right) & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{1}$$

(Takahata, 1988; Notohara, 1990; Herbots, 1997; Kozakai et al., 2016). Similarly, if

$\tau_1 < t \leq \tau_0$,

$$\mathbf{Q_2} = \begin{bmatrix} -(1+M_1) & 0 & M_1 & 1 \\ 0 & -\left(\frac{1}{b}+M_2\right) & M_2 & \frac{1}{b} \\ \frac{M_2}{2} & \frac{M_1}{2} & -\left(\frac{M_1+M_2}{2}\right) & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{2}$$

Finally, for $t > \tau_0$,

$$\mathbf{Q_3} = \begin{bmatrix} -\frac{1}{a} & 0 & 0 & \frac{1}{a} \\ 0 & -\frac{1}{a} & 0 & \frac{1}{a} \\ 0 & 0 & -\frac{1}{a} & \frac{1}{a} \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{3}$$

(Kingman, 1982a,b,c).

We denote by $\mathbf{P}(t) := \mathbf{P}(0,t)$ the transition matrix whose (i,j) entry gives the probability that the coalescent under the GIM model moves from state i at time 0 to state j at time t ($i,j \in \{1,2,3,4\}$). This transition matrix has the following form:

$$\mathbf{P}(t) = \begin{cases} e^{\mathbf{Q_1}t} & \text{for } 0 \leq t \leq \tau_1, \\ e^{\mathbf{Q_1}\tau_1}\, e^{\mathbf{Q_2}(t-\tau_1)} & \text{for } \tau_1 < t \leq \tau_0, \\ e^{\mathbf{Q_1}\tau_1}\, e^{\mathbf{Q_2}(\tau_0-\tau_1)}\, e^{\mathbf{Q_3}(t-\tau_0)} & \text{for } \tau_0 < t < \infty, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

In previous work (Costa and Wilkinson-Herbots, 2017, Appendix A, parts (ii) and (iii)) we proved[1] that if both $M_1 > 0$ and $M_2 > 0$, the matrix $\mathbf{Q_2}$ is diagonalisable and has non-positive, real eigenvalues; it was shown that three of the eigenvalues are strictly negative and one is zero. A similar argument shows that if both $M_1' > 0$ and $M_2' > 0$, the matrix $\mathbf{Q_1}$ is diagonalisable, with three strictly negative eigenvalues and one zero eigenvalue. Hence, for $M_1, M_2, M_1', M_2' > 0$, the transition matrix $\mathbf{P}(t)$ can be written

[1] On a technical note: whereas for the proof in Costa and Wilkinson-Herbots (2017), both the second and third row and the second and third column of $\mathbf{Q_2}$ were swapped round for mathematical convenience, such re-ordering of states does not affect the diagonalisability nor the eigenvalues of $\mathbf{Q_2}$.

as:

$$\mathbf{P}(t) \;=\; \begin{cases} \mathbf{G^{-1}} e^{-\mathbf{A}t}\mathbf{G} & \text{for } 0 \le t \le \tau_1, \\ \mathbf{G^{-1}} e^{-\mathbf{A}\tau_1}\mathbf{G}\,\mathbf{C^{-1}} e^{-\mathbf{B}(t-\tau_1)}\mathbf{C} & \text{for } \tau_1 < t \le \tau_0, \\ \mathbf{G^{-1}} e^{-\mathbf{A}\tau_1}\mathbf{G}\,\mathbf{C^{-1}} e^{-\mathbf{B}(\tau_0-\tau_1)}\mathbf{C}\, e^{\mathbf{Q_3}(t-\tau_0)} & \text{for } \tau_0 < t < \infty, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $\mathbf{G}$ and $\mathbf{C}$ are the matrices whose rows contain the left eigenvectors of $\mathbf{Q_1}$ and $\mathbf{Q_2}$ respectively, and where $-\mathbf{A}$ and $-\mathbf{B}$ are the corresponding diagonal matrices of non-positive, real eigenvalues. The entries in the main diagonals of $\mathbf{A}$ and $\mathbf{B}$ contain the absolute values of the eigenvalues, and are represented by the letters $\alpha_i = (\mathbf{A})_{ii}$ and $\beta_i = (\mathbf{B})_{ii}$.

If $M_1' = M_2' = 0$, then the matrix $\mathbf{Q_1}$ has eigenvalues $(-1/c_1, -1/c_2, 0, 0)$, with four linearly independent left eigenvectors given by the rows of

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

so $\mathbf{Q_1}$ is still diagonalisable. The same holds for the matrix $\mathbf{Q_2}$ when $M_1 = M_2 = 0$, with eigenvalues $(-1, -1/b, 0, 0)$. So if there is no gene flow between times τ_0 and τ_1 ago, or no gene flow between time τ_1 ago and the present, the transition matrix $\mathbf{P}(t)$ can still be decomposed as in equation (5), where in that case the matrices $\mathbf{G}$ or $\mathbf{C}$ (or both) are equal to $\mathbf{D}$.

Furthermore, for all values of M_1' and M_2', the characteristic polynomial of $\mathbf{Q_1}$, denoted $\mathcal{P}_{\mathbf{Q_1}}(x)$, is of the form $\left(x \cdot \mathcal{P}_{\mathbf{Q_1^{(r)}}}(x)\right)$, where $\mathbf{Q_1^{(r)}}$ is the 3×3 upper-left submatrix of $\mathbf{Q_1}$. So $\mathbf{Q_1}$ has a zero eigenvalue and its three remaining eigenvalues are the eigenvalues of $\mathbf{Q_1^{(r)}}$. If $M_i' = 0$ and $M_j' > 0$ $(i, j \in \{1, 2\}$ with $i \neq j)$ then, because of the resulting zero entries in $\mathbf{Q_1^{(r)}}$, it is easily seen that in this case its eigenvalues are the entries in its main diagonal. Hence the eigenvalues of $\mathbf{Q_1}$ will be $\left(-1/c_i, -(1/c_j + M_j'), -M_j'/2, 0\right)$. If these four eigenvalues are all distinct, then $\mathbf{Q_1}$ is diagonalisable. Thus, even if there is unidirectional gene flow between time τ_1 ago and the present, the transition matrix $\mathbf{P}(t)$ can still be decomposed as in equation (5), provided $\mathbf{Q_1}$ has no repeated eigenvalues.

Two comments are in order here: first, repeated eigenvalues will occur if and only if $1/c_i = M'_j/2$ or $1/c_i = 1/c_j + M'_j$; second, the set of parameter values that make these equalities true is negligible when compared to the whole parameter space, so it is very unlikely that the likelihood maximisation procedure would choose values from this set (although one should be careful to avoid using them as initial values). Similarly, if there is unidirectional gene flow between times τ_0 and τ_1 ago (i.e. $M_i = 0$ and $M_j > 0$ for $i, j \in \{1, 2\}$ with $i \neq j$), equation (5) still holds, provided $\mathbf{Q_2}$ has no repeated eigenvalues, i.e. provided the entries on the main diagonal of $\mathbf{Q_2}$ are all distinct, which is again the case for all but a negligible subset of the parameter space.

The probability that, starting in state i ($i \in \{1, 2, 3\}$), the process has reached state 4 by time t is given by the entry corresponding to the i^{th} row and 4^{th} column of the transition matrix $\mathbf{P}(t)$. This is also the cumulative distribution function of the coalescence time T_i, which we denote $F_{T_i}(t)$. If the initial state is i, and $p_{ij}^{(1)}(t)$, $p_{jl}^{(2)}(t)$ and $p_{l4}^{(3)}(t)$ denote transition probabilities of the homogeneous continuous-time Markov chains with generator matrices $\mathbf{Q_1}$, $\mathbf{Q_2}$ and $\mathbf{Q_3}$ respectively, then:

$$F_{T_i}(t) = \begin{cases} p_{i4}^{(1)}(t) & \text{for } 0 \leq t \leq \tau_1, \\ \sum_{j=1}^{4} p_{ij}^{(1)}(\tau_1)\, p_{j4}^{(2)}(t - \tau_1) & \text{for } \tau_1 < t \leq \tau_0, \\ \sum_{j=1}^{4} p_{ij}^{(1)}(\tau_1) \sum_{l=1}^{4} p_{jl}^{(2)}(\tau_0 - \tau_1)\, p_{l4}^{(3)}(t - \tau_0) & \text{for } \tau_0 < t < \infty, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Denoting by H_{mn} the (m, n) entry of a matrix $\mathbf{H}$, and by H_{mn}^{-1} the (m, n) entry of the matrix $\mathbf{H^{-1}}$, we have for $t \geq 0$ that $p_{ij}^{(1)}(t) = \sum_{k=1}^{4} G_{ik}^{-1} G_{kj}\, e^{-\alpha_k t}$ and $p_{jl}^{(2)}(t) = \sum_{k=1}^{4} C_{jk}^{-1} C_{kl}\, e^{-\beta_k t}$. Furthermore, $p_{l4}^{(3)}(t) = 1 - e^{-\frac{1}{a}t}$ for $l = 1, 2, 3$, as the absorption time from state l into state 4 of the homogeneous continuous-time Markov Chain generated by $\mathbf{Q_3}$ is exponentially distributed with mean a. Using that $p_{44}^{(1)}(t) = p_{44}^{(2)}(t) = p_{44}^{(3)}(t) = 1$ for all $t \geq 0$, differentiating equation (6) gives the following expression for the probability

density function of T_i:

$$f_{T_i}(t) = \begin{cases} f_i^{(1)}(t) & \text{for } 0 \le t \le \tau_1, \\ \sum_{j=1}^{3} p_{ij}^{(1)}(\tau_1)\, f_j^{(2)}(t-\tau_1) & \text{for } \tau_1 < t \le \tau_0, \\ \sum_{j=1}^{3} p_{ij}^{(1)}(\tau_1) \sum_{l=1}^{3} p_{jl}^{(2)}(\tau_0-\tau_1)\, f_l^{(3)}(t-\tau_0) & \text{for } \tau_0 < t < \infty, \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where for $t > 0$, $f_i^{(1)}(t) = -\sum_{k=1}^{4} \alpha_k\, G_{ik}^{-1} G_{k4}\, e^{-\alpha_k t}$, $f_j^{(2)}(t) = -\sum_{k=1}^{4} \beta_k\, C_{jk}^{-1} C_{k4}\, e^{-\beta_k t}$ and $f_l^{(3)}(t) = \frac{1}{a} e^{-\frac{1}{a}t}$ are the probability density functions of the absorption times into state 4 of the three homogeneous continuous-time Markov Chains generated by $\mathbf{Q_1}$, $\mathbf{Q_2}$ and $\mathbf{Q_3}$, starting in states i, j and l, respectively ($i, j, l \in \{1, 2, 3\}$).

2.2 The distribution of the number of pairwise nucleotide differences

We assume that selectively neutral mutations occur according to the infinite sites model of Watterson (1975), in which the locus under consideration consists of an infinite sequence of nucleotide sites and no two mutations ever occur at the same site; in each generation, the number of mutations occurring in a DNA sequence at this locus is Poisson distributed with mean μ, and it is assumed that mutations occur independently in different DNA sequences and in different generations. In the coalescent approximation, measuring time in units of $2N$ generations, mutations then accumulate on each ancestral lineage according to a Poisson process of rate $\theta/2$, where $\theta = 4N\mu$ is the 'scaled' mutation rate. Given the coalescence time T_i of two DNA sequences, the number of nucleotide differences between them, denoted by S_i, is simply the total number of mutations that have accumulated on their ancestral lineages since their most recent common ancestor, and hence is Poisson distributed with mean θT_i; as before, the subscript i refers to the initial state of the coalescent process, corresponding to the sampling locations of the pair of DNA sequences ($i \in \{1, 2, 3\}$). Denoting $g_s(t) := \frac{(\theta t)^s}{s!} e^{-\theta t}$ and using equation (7), the probability of s nucleotide differences between the two sequences can be written as

follows, for $s = 0, 1, 2, \ldots$:

$$\begin{aligned}
\mathrm{P}(S_i = s) &= \mathrm{E}[g_s(T_i)] \\
&= \int_0^{\tau_1} g_s(t)\, f_i^{(1)}(t)\, \mathrm{d}t + \sum_{j=1}^{3} p_{ij}^{(1)}(\tau_1) \int_{\tau_1}^{\tau_0} g_s(t)\, f_j^{(2)}(t-\tau_1)\, \mathrm{d}t \\
&\quad + \sum_{j=1}^{3} p_{ij}^{(1)}(\tau_1) \sum_{l=1}^{3} p_{jl}^{(2)}(\tau_0-\tau_1) \int_{\tau_0}^{\infty} g_s(t)\, f_l^{(3)}(t-\tau_0)\, \mathrm{d}t \quad .
\end{aligned}$$

Changing the limits of integration, and using the expressions for $f_i^{(1)}(t)$, $f_j^{(2)}(t)$ and $f_l^{(3)}(t)$ given in the previous section, the above equation becomes:

$$\begin{aligned}
\mathrm{P}\left(S_i = s\right) &= \int_0^{\tau_1} g_s(t)\, f_i^{(1)}(t)\, \mathrm{d}t + \sum_{j=1}^{3} p_{ij}^{(1)}(\tau_1) \int_0^{\tau_0-\tau_1} g_s(\tau_1+t)\, f_j^{(2)}(t)\, \mathrm{d}t \\
&\quad + \sum_{j=1}^{3} p_{ij}^{(1)}(\tau_1) \sum_{l=1}^{3} p_{jl}^{(2)}(\tau_0-\tau_1) \int_0^{\infty} g_s(\tau_0+t)\, f_l^{(3)}(t)\, \mathrm{d}t \\
&= -\sum_{k=1}^{4} \alpha_k\, G_{ik}^{-1} G_{k4} \int_0^{\tau_1} g_s(t)\, e^{-\alpha_k t}\, \mathrm{d}t \\
&\quad - \sum_{j=1}^{3} p_{ij}^{(1)}(\tau_1) \sum_{k=1}^{4} \beta_k\, C_{jk}^{-1} C_{k4} \int_0^{\tau_0-\tau_1} g_s(\tau_1+t)\, e^{-\beta_k t}\, \mathrm{d}t \\
&\quad + \frac{1}{a} \sum_{j=1}^{3} p_{ij}^{(1)}(\tau_1) \sum_{l=1}^{3} p_{jl}^{(2)}(\tau_0-\tau_1) \int_0^{\infty} g_s(\tau_0+t)\, e^{-\frac{1}{a}t}\, \mathrm{d}t \quad .
\end{aligned}$$

Recall that some eigenvalues of $\mathbf{Q_1}$ and $\mathbf{Q_2}$ are equal to zero, i.e. some of the α_k and β_k in the above expression are zero. For those α_k and β_k that are strictly positive, we let W_k and Y_k denote exponentially distributed random variables with rates α_k and β_k respectively, and we denote by X an exponentially distributed random variable with

rate $1/a$. The equation above can then be written as:

$$
\begin{aligned}
&\mathrm{P}\left(S_i = s\right) \\
&= -\sum_{k:\alpha_k>0} G_{ik}^{-1} G_{k4}\, \mathrm{E}[g_s(W_k)|W_k \le \tau_1]\, \mathrm{P}(W_k \le \tau_1) \\
&\quad -\sum_{j=1}^{3} p_{ij}^{(1)}(\tau_1) \sum_{k:\beta_k>0} C_{jk}^{-1} C_{k4}\, \mathrm{E}[g_s(\tau_1 + Y_k)|\tau_1 + Y_k \le \tau_0]\, \mathrm{P}(\tau_1 + Y_k \le \tau_0) \\
&\quad +\sum_{j=1}^{3} p_{ij}^{(1)}(\tau_1) \sum_{l=1}^{3} p_{jl}^{(2)}(\tau_0 - \tau_1)\, \mathrm{E}[g_s(\tau_0 + X)] \\
&= -\sum_{k:\alpha_k>0} G_{ik}^{-1} G_{k4} \left\{\mathrm{E}[g_s(W_k)] - \mathrm{E}[g_s(W_k)|W_k > \tau_1]\mathrm{P}(W_k > \tau_1)\right\} \\
&\quad -\sum_{j=1}^{3} p_{ij}^{(1)}(\tau_1) \sum_{k:\beta_k>0} C_{jk}^{-1} C_{k4} \left\{\mathrm{E}[g_s(\tau_1 + Y_k)] - \mathrm{E}[g_s(\tau_1 + Y_k)|\tau_1 + Y_k > \tau_0]\, \mathrm{P}(\tau_1 + Y_k > \tau_0)\right\} \\
&\quad +\sum_{j=1}^{3} p_{ij}^{(1)}(\tau_1) \sum_{l=1}^{3} p_{jl}^{(2)}(\tau_0 - \tau_1)\, \mathrm{E}[g_s(\tau_0 + X)] \quad .
\end{aligned}
$$

Finally, making use of the lack of memory property of the exponential distribution, we obtain:

$$
\begin{aligned}
\mathrm{P}(S_i = s) \quad = \quad & -\sum_{k:\alpha_k>0} G_{ik}^{-1} G_{k4} \left\{\mathrm{E}[g_s(W_k)] - \mathrm{E}[g_s(\tau_1 + W_k)]\, e^{-\alpha_k \tau_1}\right\} \\
& -\sum_{j=1}^{3} p_{ij}^{(1)}(\tau_1) \sum_{k:\beta_k>0} C_{jk}^{-1} C_{k4} \left\{\mathrm{E}[g_s(\tau_1 + Y_k)] - \mathrm{E}[g_s(\tau_0 + Y_k)]\, e^{-\beta_k(\tau_0 - \tau_1)}\right\} \\
& +\sum_{j=1}^{3} p_{ij}^{(1)}(\tau_1) \sum_{l=1}^{3} p_{jl}^{(2)}(\tau_0 - \tau_1)\, \mathrm{E}[g_s(\tau_0 + X)] \quad .
\end{aligned}
\tag{8}
$$

Thus the probability that two DNA sequences sampled from locations given by initial state $i \in \{1, 2, 3\}$ differ at s nucleotide sites (and similarly, the expectation of any other function of the coalescence time T_i) can be obtained from results for Exponential and shifted Exponential random variables. In particular, for an exponentially distributed random variable U with rate parameter λ, $\mathrm{E}[g_s(U)] = \mathrm{E}\left[\frac{(\theta U)^s}{s!}\, e^{-\theta U}\right]$ is the probability that s events occur in a Poisson Process of rate θ during a time span of length U; this can be written as a probability from a geometric distribution:

$$
\mathrm{E}[g_s(U)] = \left(\frac{\theta}{\lambda + \theta}\right)^s \frac{\lambda}{\lambda + \theta} \qquad \text{for } s = 0, 1, 2, \ldots \tag{9}
$$

(Watterson, 1975). Similarly, for any $\tau > 0$, $\mathrm{E}[g_s(\tau+U)]$ is the probability that s events occur in a Poisson Process of rate θ during a time span of length $\tau + U$, which is given by

$$\mathrm{E}[g_s(\tau+U)] = e^{-\theta\tau}\frac{\lambda\,\theta^s}{(\lambda+\theta)^{s+1}}\sum_{l=0}^{s}\frac{(\lambda+\theta)^l\,\tau^l}{l!} \qquad \text{for } s = 0,1,2,\ldots \tag{10}$$

(Takahata et al., 1995; see also Wilkinson-Herbots, 2008). Substituting (9) and (10) into equation (8) then gives the following result for the probability distribution of S_i, the number of nucleotide differences between two DNA sequences sampled from locations given by initial state $i \in \{1,2,3\}$:

$$\begin{aligned}
\mathrm{P}(S_i = s) \;=\; & -\sum_{k:\alpha_k>0} G^{-1}_{ik}G_{k4}\,\frac{\alpha_k\,\theta^s}{(\alpha_k+\theta)^{s+1}}\left(1 - e^{-(\alpha_k+\theta)\tau_1}\sum_{l=0}^{s}\frac{(\alpha_k+\theta)^l\,{\tau_1}^l}{l!}\right) \\
& -\sum_{j=1}^{3} p^{(1)}_{ij}(\tau_1)\sum_{k:\beta_k>0} C^{-1}_{jk}C_{k4}\,\frac{\beta_k\,\theta^s}{(\beta_k+\theta)^{s+1}}\left(e^{-\theta\tau_1}\sum_{l=0}^{s}\frac{(\beta_k+\theta)^l\,{\tau_1}^l}{l!}\right. \\
& \qquad\qquad \left. -e^{-(\beta_k+\theta)\tau_0+\beta_k\tau_1}\sum_{l=0}^{s}\frac{(\beta_k+\theta)^l\,\tau_0^l}{l!}\right) \\
& +\left(\sum_{j=1}^{3} p^{(1)}_{ij}(\tau_1)\sum_{l=1}^{3} p^{(2)}_{jl}(\tau_0-\tau_1)\right) e^{-\theta\tau_0}\,\frac{(a\theta)^s}{(1+a\theta)^{s+1}}\sum_{l=0}^{s}\frac{\left(\frac{1}{a}+\theta\right)^l\,\tau_0^l}{l!}
\end{aligned} \tag{11}$$

for $s = 0,1,2,\ldots$ Recalling also that $p^{(1)}_{ij}(\tau_1) = \sum_{k=1}^{4} G^{-1}_{ik}G_{kj}\,e^{-\alpha_k\tau_1}$ and $p^{(2)}_{jl}(\tau_0-\tau_1) = \sum_{k=1}^{4} C^{-1}_{jk}C_{kl}\,e^{-\beta_k(\tau_0-\tau_1)}$, the above probability can easily be computed for any parameter values, using standard numerical procedures to compute the eigenvalues and eigenvectors of the matrices $\mathbf{Q_1}$ and $\mathbf{Q_2}$, except for a negligible subset of parameter values for which the matrices $\mathbf{Q_1}$ and/or $\mathbf{Q_2}$ are not diagonalisable (see Subsection 2.1).

If $M_1' = M_2' = 0$, then equation (11) reduces to the corresponding results for the 'isolation with initial migration' (IIM) model, given by equations (11) and (12) in Costa and Wilkinson-Herbots (2017). If, in addition, $M_1 = M_2$ and $b = 1$ (an IIM model with symmetric migration and equal population sizes during the migration period), then explicit expressions are available for the eigenvalues of the matrix $\mathbf{Q_2}$, and equation (11) simplifies to the fully explicit expressions given in Wilkinson-Herbots (2012), equations (18) and (29).

2.3 The likelihood of a multilocus data set

Recall that, for the purposes of this paper, an observation consists of the number of nucleotide differences between two DNA sequences at a given locus. To jointly estimate all the parameters of the GIM model, our method requires a large set of observations, all at different loci, from each of the three possible initial states: both sequences sampled from descendant population 1 (state 1), both sequences sampled from descendant population 2 (state 2), or one sequence sampled from each of the two descendant populations (state 3). To compute the likelihood of such a data set, we will assume that all observations are independent, so our data should include no more than one observation (i.e. pair of sequences) per locus and there should be free recombination between loci, i.e. all loci should be sufficiently far apart.

Let $\boldsymbol{\rho}$ be the vector of parameters of the coalescent under the GIM model, i.e.

$$\boldsymbol{\rho} = (a, b, c_1, c_2, \tau_1, \tau_0, M_1, M_2, M_1', M_2') \quad .$$

Denote by J_i the number of loci at which the two sampled DNA sequences are from locations corresponding to initial state i ($i = 1, 2, 3$). We redefine θ to denote the *average* scaled mutation rate over all loci in the combined data set made up of the observations from all three initial states. For $i = 1, 2, 3$ and $j = 1, \ldots, J_i$, denote by $\theta_{ij} = 4N\mu_{ij}$ the scaled mutation rate of the jth locus associated with initial state i, where μ_{ij} is the mutation rate per sequence per generation at that locus, and denote by $r_{ij} = \frac{\theta_{ij}}{\theta}$ the *relative* mutation rate of that locus, so that $\theta_{ij} = r_{ij}\theta$. Assuming that the relative mutation rates are known, and denoting by s_{ij} the observation at the jth locus associated with initial state i (i.e. the number of nucleotide differences between the two DNA sequences sampled at that locus), the likelihood of the data set $\boldsymbol{s} = (s_{ij})_{i=1,2,3;j=1,\ldots,J_i}$ can be written as

$$L\left(\boldsymbol{\rho}, \theta; \mathbf{s}\right) = \prod_{i=1}^{3} \prod_{j=1}^{J_i} L(\boldsymbol{\rho}, \theta; s_{ij})$$

where $L(\boldsymbol{\rho}, \theta; s_{ij})$, the likelihood of the observation s_{ij}, is the probability that the two DNA sequences sampled at the jth locus associated with initial state i differ at s_{ij} nucleotide sites and is given by equation (11) with θ replaced by $\theta_{ij} = r_{ij}\theta$, and s replaced by s_{ij}.

In our maximum-likelihood method, the relative mutation rates r_{ij} are treated as known constants. In practice, however, the relative mutation rates at the different loci are usually estimated using outgroup sequences (for example, Yang, 2002; Wang and Hey, 2010; Lohse et al., 2011; Costa and Wilkinson-Herbots, 2017). It should be noted that standard errors and confidence intervals obtained with our method do not account for uncertainty about the relative mutation rates.

To increase the robustness and performance of the likelihood maximisation procedure, our computer implementation uses the following reparameterisations:

$$\begin{aligned} &\theta_0 = a\theta,\ \theta_1 = \theta,\ \theta_2 = b\theta,\ \theta'_1 = c_1\theta,\ \theta'_2 = c_2\theta, \\ &T_1 = \theta\tau_1,\ V = \theta(\tau_0 - \tau_1), \\ &M_1{}^* = M_1,\ M_2{}^* = bM_2,\ M'_1{}^* = c_1M'_1,\ M'_2{}^* = c_2M'_2 \end{aligned} \tag{12}$$

(similar to the choice of parameters in, for example, Hey and Nielsen, 2004; Zhu and Yang, 2012; Costa and Wilkinson-Herbots, 2017); see also Figure 2. With this reparameterisation, θ_0 , θ_1 , θ_2 , θ'_1 and θ'_2 are the 'population size parameters' of, respectively, the ancestral population, descendant populations 1 and 2 between times τ_0 and τ_1 ago, and descendant populations 1 and 2 between time τ_1 ago and the present. Note that for each population, $\theta_i = 4N_i\mu$, where $2N_i$ is the size (the number of DNA sequences at any locus) of the population concerned and μ is the mutation rate per sequence per generation averaged over all the loci in the data set; similarly for θ'_i. The 'time' parameters T_1 and V represent, respectively, the durations of the most recent stage of the model (i.e. from time τ_1 ago until the present) and the intermediate stage of the

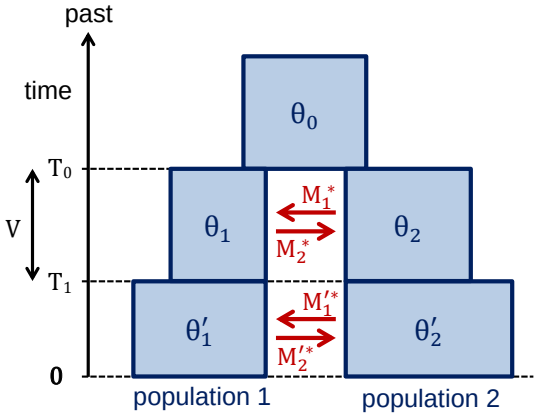

Figure 2: The reparameterised GIM model. The direction of migration shown is from a forward-in-time perspective.

model (between times τ_0 and τ_1 ago), but these durations are now measured by twice the expected number of mutations per DNA sequence during the period concerned. For $i = 1, 2$, the migration parameter M_i^* is, from a forward in time perspective, twice the number of immigrant DNA sequences into descendant population i per generation between times τ_0 and τ_1 ago (equivalently, if we work backward in time, it is twice the number of DNA sequences that migrate from population i per generation between times τ_1 and τ_0 ago). Similarly, $M_i'^*$ is twice the number of immigrant DNA sequences into descendant population i per generation between time τ_1 ago and the present (or, from a backward in time perspective, twice the number of emigrant DNA sequences from descendant population i per generation during this period). Our computer code for fitting the full GIM model to the data obtains ML estimates jointly for the 11 parameters θ_0 , θ_1 , θ_2 , θ_1' , θ_2' , T_1 , V, M_1^*, M_2^*, $M_1'^*$, $M_2'^*$. These estimates can readily be converted to ML estimates of the original model parameters if required.

Our computer implementation also allows the three simpler models illustrated in Figure 1 to be fitted to the data, using the same reparameterisation as above: (a) a model of complete isolation that allows for a change of the descendant population sizes ($M_1^* = M_2^* = M_1'^* = M_2'^* = 0$, leaving 7 parameters to be estimated); (b) the 'isolation with initial migration' model ($M_1'^* = M_2'^* = 0$, leaving 9 parameters to be estimated; see also Costa and Wilkinson-Herbots, 2017); and (c) a model of secondary contact ($M_1^* = M_2^* = 0$, again leaving 9 parameters to be estimated). For each of these models, the computation of the likelihood uses an appropriately simplified version of equation (11). Versions of the GIM model involving unidirectional gene flow during the most recent and/or the intermediate stage of the model can also readily be implemented, but are not further considered in this paper.

2.4 Model comparison

For any particular data set, a straightforward way to compare the fit of the GIM model and that of simpler models nested within it is by using Akaike's Information Criterion, AIC, which was designed to compare competing models with different numbers of parameters. For each model,

$$\text{AIC} = -2 \ln \hat{L} + 2k$$

(Akaike, 1972, 1974), where $\hat{L}$ is the maximised likelihood of the model, given the data, and k is the number of free parameters in the model. Thus a larger maximised likelihood leads to a lower AIC value, subject to a penalty for each additional model parameter. The 'best' model amongst the competing models considered is that with the smallest AIC value – this model (with the maximum-likelihood estimates of its parameters) is called the 'Minimum AIC Estimate' (MAICE).

An alternative approach is to perform a series of likelihood ratio tests for pairs of nested models. For example, if we wish to compare the four models depicted in Figure 1, we can start by assuming the simplest of these four models (i.e. that with the smallest number of parameters) as the null hypothesis: the isolation model (Figure 1a). We can then proceed by performing two likelihood ratio tests: one where we test the isolation model (H_0) against the alternative hypothesis of the IIM model (which we will denote by $H_{1,1}$); and one where we test the isolation model (H_0) against the model of secondary contact (denoted by $H_{1,2}$). Since we are performing two tests, a procedure for controlling either the family-wise error rate (for example, a Bonferroni correction), or the false discovery rate, should be applied. Depending on the results of these two significance tests, we then proceed as follows. If neither test gives a significant result, then we retain the isolation model as our 'best' model and conclude that there is no significant evidence of gene flow at any time between time τ_0 ago and the present. If the test of H_0 against $H_{1,1}$ is significant, but the test of H_0 against $H_{1,2}$ is not, then we reject the isolation model in favour of the IIM model; in this case we then proceed by assuming the IIM model to be our new null hypothesis and testing this against the full GIM model as our new alternative hypothesis (no further correction for multiple testing is required, as this third LR test is only performed if the first LR test gave a significant result). Similarly, if the test of H_0 against $H_{1,2}$ is significant, but the test of H_0 against $H_{1,1}$ is not, then we reject the isolation model in favour of the model of secondary contact; we then proceed by taking the latter model to be our new null hypothesis and testing it against the full GIM model. If the first two LR tests (H_0 against $H_{1,1}$ and H_0 against $H_{1,2}$) are *both* significant, then we will reject the isolation model in favour of the alternative model that has the highest likelihood: $H_{1,1}$ or $H_{1,2}$ i.e. the IIM model or the model of secondary contact; we then use that model as our new null hypothesis and test it against the full GIM model.

In each of the Likelihood Ratio tests described above, the alternative model has two more free parameters than the null model, namely the migration rates in both directions in either the most recent or the intermediate stage of the model. For each test, the null model corresponds to the two migration rates concerned being zero, i.e. a parameter value on the boundary of the parameter space. The null distribution of the LRT statistic

$$\Lambda = 2 \times \Big(\ln \hat{L}(\text{alternative model}) - \ln \hat{L}(\text{null model}) \Big)$$

is therefore not χ^2_2, but a mixture of χ^2_ν distributions for $\nu = 0, 1, 2$ (Self and Liang, 1987; Silvapulle and Sen, 2005); using χ^2_2 instead of the correct null distribution is conservative. In the mixture of χ^2_ν distributions, computation of the weights of χ^2_0 and χ^2_2 is not straightforward, but the weight of χ^2_1 is known to be $\frac{1}{2}$ (Self and Liang, 1987; Silvapulle and Sen, 2005). The use of the mixture $\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2$ instead of the correct null distribution is therefore also conservative, and results in a smaller loss of power compared to the use of χ^2_2. In the literature, the mixture $\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$ has also been used as the null distribution when testing for gene flow in the simpler isolation-with-migration model (Wang and Hey, 2010). The use of this latter mixture results in a smaller loss of power compared to the use of the mixture $\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2$ suggested above; however, to the best of our knowledge it is yet to be established whether the use of $\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$ instead of the precise asymptotic null distribution in the context of testing for gene flow in the IM or GIM models is always conservative (i.e. whether the weight of χ^2_2 in the mixture that makes up the correct asymptotic null distribution is $\leq \frac{1}{4}$).

Both methods of model selection described above, using AIC scores or likelihood ratio tests, will be examined for simulated data in the next Section. In the case of likelihood ratio tests, we will also consider the performance of all three 'null distributions' suggested in the previous paragraph.

3 Simulation results

To examine the accuracy of ML estimates of the parameters of the GIM model obtained with our code, and to examine whether our method makes it possible to distinguish between the different models considered in Figure 1, we simulated 200 data sets from each of the following five scenarios:

(i) a GIM model with decreasing gene flow (i.e. where the migration rates in the most recent stage of the model are lower than in the intermediate stage);

(ii) a GIM model with increasing gene flow (i.e. where the migration rates in the most recent stage of the model are higher than in the intermediate stage);

(iii) a GIM model with gene flow decreasing to zero, i.e. an isolation-with-initial-migration model;

(iv) a GIM model with zero migration rates in the intermediate stage of the model, i.e. a model of secondary contact;

(v) a GIM model with all migration rates equal to zero, i.e. a model of complete isolation.

Each simulated data set consists of the numbers of nucleotide differences between one pair of DNA sequences sampled at each of 40,000 independent loci: for 10,000 loci, two DNA sequences were sampled both from descendant population 1; for 10,000 loci, two DNA sequences were sampled from descendant population 2; and for 20,000 loci, one DNA sequence was sampled from each of the two descendant populations. The relative mutation rates of the 40,000 loci were simulated from a Gamma$(10, 10)$ distribution. The R function used to generate the data is available at https://github.com/Costa-and-Wilkinson-Herbots/GIM .

The 'true' values of the population size parameters and time parameters assumed for the simulations were as follows: $\theta_0 = 3,\ \theta_1 = 2,\ \theta_2 = 4,\ \theta_1' = 3,\ \theta_2' = 6,\ T_1 = 4,\ V = 4.$ These parameter values were based on a hypothetical scenario where the sampled loci have an average length of 500 nucleotide sites, with an average mutation rate of 10^{-9} per site per generation, and with population sizes of the order of a million individuals (for example, $\theta_1 = 2$ corresponds in this case to a population size of 2 million DNA sequences, or 1 million diploid individuals); this order of magnitude of the mutation rate and effective population sizes may, for example, be broadly realistic for some species of *Drosophila* (Wang and Hey, 2010; Keightley et al., 2014). The durations of the intermediate and the most recent stage of the model correspond to 2 expected mutations per DNA sequence during each of these time periods, which in this hypothetical scenario would equate to 4 million generations each.

For scenarios (i) to (v) listed above, the migration parameters were assumed to have the following 'true' values:

(i) $M_1^* = 0.2$ and $M_2^* = 0.4$, decreasing by a factor of five to $M_1'^* = 0.04$ and $M_2'^* = 0.08$;

(ii) $M_1^* = 0.04$ and $M_2^* = 0.08$, increasing five-fold to $M_1'^* = 0.2$ and $M_2'^* = 0.4$;

(iii) $M_1^* = 0.2$ and $M_2^* = 0.4$, decreasing to $M_1'^* = M_2'^* = 0$;

(iv) $M_1^* = M_2^* = 0$, increasing to $M_1'^* = 0.2$ and $M_2'^* = 0.4$;

(v) $M_1^* = M_2^* = M_1'^* = M_2'^* = 0$;

recall that M_i^* and $M_i'^*$ are twice the numbers of immigrant DNA sequences into descendant population i per generation during, respectively, the intermediate and the most recent stage of the model.

To investigate the accuracy of the ML estimates obtained with our GIM code, we first fitted a GIM model to each simulated data set. Thus, for each simulated data set, ML estimates of the 11 parameters of the GIM model were obtained, while the relative mutation rates of the 40,000 loci were treated as known constants. Boxplots of the 200 sets of parameter estimates obtained for each of the five scenarios are shown in Figure 3. In each scenario, it is seen that for all 11 parameters, the median estimate obtained (the bold line in each boxplot) is close to the true parameter value (indicated by a red cross in each boxplot); however, in scenario (ii) (GIM model with increasing gene flow), the median of the estimates obtained for M_1^* was 0, whereas the true value of this parameter assumed in the simulations was very small but non-zero ($M_1^* = 0.04$). The plots also suggest that, while the population size parameters of the descendant populations during the most recent stage of the model and of the ancestral population can be estimated with high precision, the estimates of the population size parameters during the intermediate stage of the model display considerably more variability. The estimates of the migration parameters are also quite variable, again particularly so for the intermediate stage of the model. Table 1 lists, for each of the five scenarios and for each of the 11 parameters, the mean and standard deviation of the estimates obtained, as well as the relative bias (i.e. the bias divided by the true parameter value). It is seen that for most parameters and scenarios, the relative bias is very small: less than 1% in most cases, and less than 10%

in all but three cases. These three exceptions all concern the intermediate stage of the model in scenarios of increasing gene flow: the estimates of the migration parameters M_1^* and M_2^* in scenario (ii), and the estimates of one of the population size parameters, θ_1, in scenario (iv); in these cases the relative bias is larger, but not excessive.

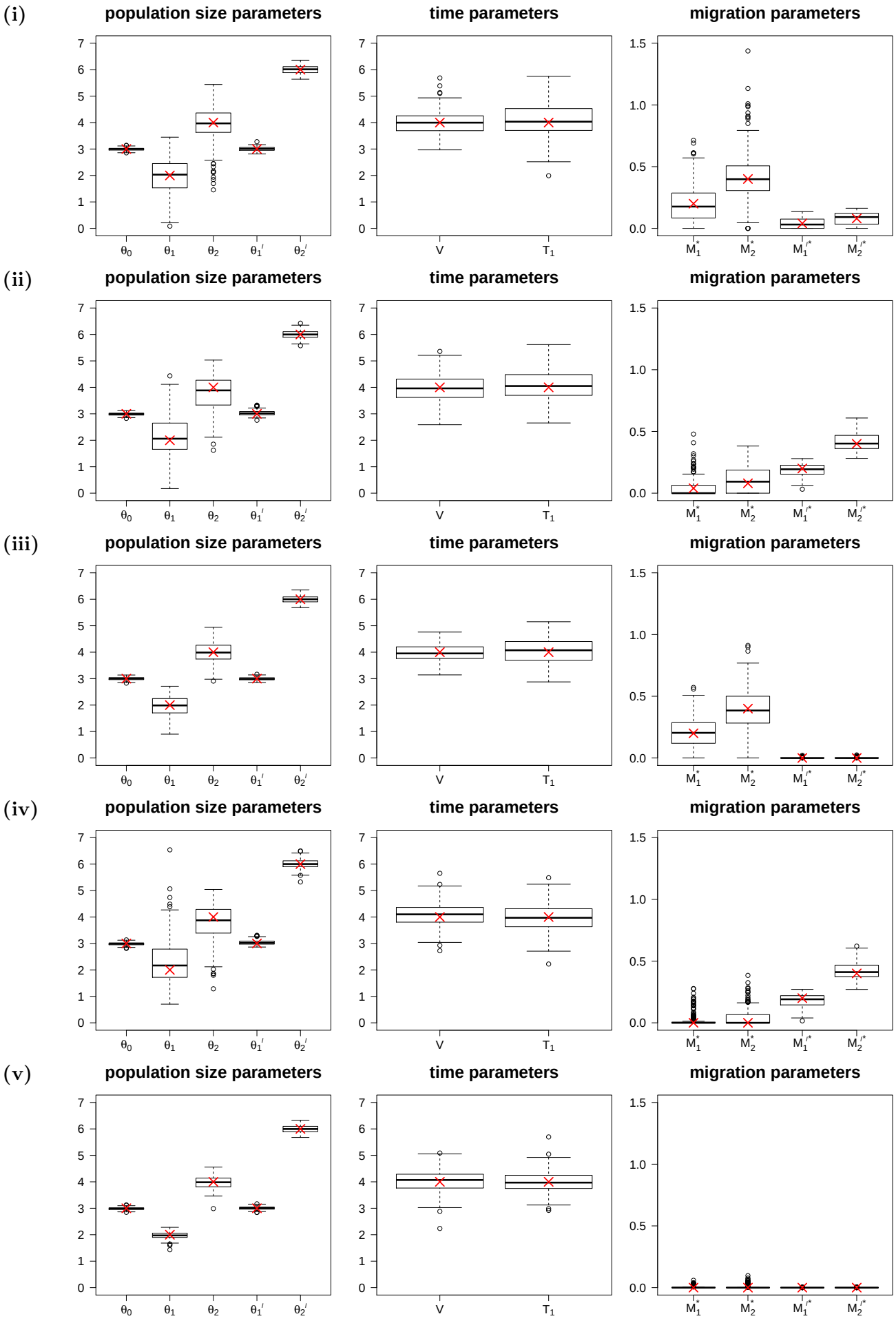

Figure 3: Box plots of the ML estimates of the parameters of the GIM model obtained for 200 simulated data sets under each of scenarios (i) to (v). The 'true' value of each parameter is indicated by a red cross.

To examine whether our method makes it possible to distinguish between the four models considered in Figure 1, we fitted all four models (GIM, IIM, secondary contact, isolation) to each of our 1,000 simulated data sets (the 200 data sets simulated from each of the five scenarios listed at the start of this Section). For each simulated data set we applied the model selection procedures described in Subsection 2.4 to select the best-fitting model: we did this using either the AIC criterion, or a sequence of Likelihood Ratio tests at a significance level of 5% (using a Bonferroni correction where appropriate); for the approach using Likelihood Ratio tests, as the precise asymptotic null distribution is not easy to compute, results were obtained using each of the three distributions suggested in Subsection 2.4: χ^2_2 , $\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2$ or $\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$. Table 2 shows, for each of our five scenarios, for what proportion of the 200 simulated data sets each model was selected as the best-fitting model; the proportion of simulated data sets for which the correct model was selected is highlighted in bold in each case. It is seen that for scenarios (i), (iii), (iv) and (v), the correct model was selected for nearly all of the simulated data sets (96% or better), for both methods (AIC or LRT) and for all three 'null distributions' considered. However for most of the data sets simulated from scenario (ii), i.e. a GIM model with gene flow *increasing* from $(M_1^*, M_2^*) = (0.04, 0.08)$ to $(M_1'^*, M_2'^*) = (0.2, 0.4)$, the model of secondary contact was selected as the best-fitting model instead of the 'true' GIM model; so whilst our method correctly inferred an *increase* of gene flow for all these data sets, its power to detect the small amount of gene flow that occurred in the intermediate stage of the model was low. This starkly contrasts with our results for scenario (i), i.e. a GIM model with gene flow *decreasing* from $(M_1^*, M_2^*) = (0.2, 0.4)$ to $(M_1'^*, M_2'^*) = (0.04, 0.08)$: for this scenario, our method correctly identified the GIM model as the best-fitting model for 99.5% of the simulated data sets, demonstrating very high power to detect the small level of gene flow in the most recent stage of the model. It should perhaps also be noted that for none of the 800 data sets simulated from scenarios with gene flow (scenarios (i) to (iv)), the isolation model was selected as the best-fitting model, i.e. the overall power of our method to detect that gene flow had occurred at some point in the past (i.e. a departure from the isolation model) was very high. However, our method's ability to apportion the inferred gene flow accurately to the two different time periods appears to be limited in the case of increasing gene flow.

Whilst the use of χ^2_2 or $\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2$ instead of the precise asymptotic null distribution

Table 1:
Mean, standard deviation and relative bias of the ML estimates of the parameters of the GIM model

		θ_0	θ_1	θ_2	θ'_1	θ'_2	T_1	V	$M_1{}^*$	$M_2{}^*$	$M'_1{}^*$	$M'_2{}^*$
scenario (i)	true parameter value	3	2	4	3	6	4	4	0.2	0.4	0.04	0.08
	mean estimate (standard deviation)	2.9923 (0.0572)	1.9864 (0.6569)	3.9184 (0.6608)	3.0047 (0.0791)	6.0080 (0.1623)	4.0884 (0.5997)	3.9884 (0.4365)	0.2004 (0.1565)	0.4247 (0.2034)	0.0417 (0.0392)	0.0794 (0.0491)
	relative bias	-0.0026	-0.0068	-0.0204	0.0016	0.0013	0.0221	-0.0029	0.0018	0.0618	0.0430	-0.0072
scenario (ii)	true parameter value	3	2	4	3	6	4	4	0.04	0.08	0.2	0.4
	mean estimate (standard deviation)	2.9890 (0.0531)	2.1823 (0.7526)	3.7696 (0.6625)	3.0262 (0.0947)	5.9992 (0.1625)	4.0681 (0.6038)	3.9740 (0.5610)	0.0469 (0.0808)	0.1072 (0.1011)	0.1872 (0.0509)	0.4124 (0.0682)
	relative bias	-0.0037	0.0912	-0.0576	0.0087	-0.0001	0.0170	-0.0065	0.1714	0.3398	-0.0640	0.0310
scenario (iii)	true parameter value	3	2	4	3	6	4	4	0.2	0.4	0	0
	mean estimate (standard deviation)	3.0001 (0.0567)	1.9509 (0.3971)	3.9890 (0.3810)	2.9941 (0.0567)	6.0031 (0.1322)	4.0452 (0.4791)	3.9843 (0.3294)	0.2081 (0.1256)	0.3948 (0.1615)	0.0012 (0.0031)	0.0013 (0.0036)
	relative bias	0.0000	-0.0246	-0.0027	-0.0020	0.0005	0.0113	-0.0039	0.0405	-0.0130	–	–
scenario (iv)	true parameter value	3	2	4	3	6	4	4	0	0	0.2	0.4
	mean estimate (standard deviation)	2.9858 (0.0533)	2.3197 (0.8562)	3.7841 (0.6959)	3.0371 (0.0904)	6.0014 (0.1772)	3.9821 (0.4955)	4.0765 (0.4492)	0.0224 (0.0538)	0.0432 (0.0738)	0.1802 (0.0542)	0.4223 (0.0718)
	relative bias	-0.0047	0.1598	-0.0540	0.0124	0.0002	-0.0045	0.0191	–	–	-0.0990	0.0557
scenario (v)	true parameter value	3	2	4	3	6	4	4	0	0	0	0
	mean estimate (standard deviation)	2.9876 (0.0494)	1.9715 (0.1300)	3.9788 (0.2272)	3.0064 (0.0568)	5.9985 (0.1394)	4.0077 (0.4016)	4.0211 (0.4124)	0.0046 (0.0099)	0.0065 (0.0156)	0.0001 (0.0006)	0.0003 (0.0009)
	relative bias	-0.0041	-0.0143	-0.0053	0.0021	-0.0002	0.0019	0.0053	–	–	–	–

Summary statistics based on 200 simulated data sets for each scenario. Each simulated data set consists of the number of nucleotide differences between one pair of DNA sequences at each of 40,000 different loci (two sequences from population 1 at 10,000 loci; two sequences from population 2 at 10,000 loci; one sequence from each population at 20,000 loci). When the true value of a parameter is 0, the relative bias (i.e. the bias divided by the true parameter value) is undefined, indicated by a '–' symbol in the table.

in our sequence of LR tests is known to be conservative, the results in Table 2 suggest that the use of $\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$ also appears to be conservative: using the latter distribution instead of the correct null distribution, and an overall significance level of 5%, the isolation model was falsely rejected (resulting in a type 1 error) for only 1.5% of

Table 2:
Model selection for simulated data: Results

simulation scenario	true model	method	best-fitting model			
			GIM	**IIM**	**secondary contact**	**isolation**
(i)	**GIM**	AIC	**99.5 %**	0 %	0.5 %	0 %
		LRT (χ^2_2)	**99.5 %**	0 %	0.5 %	0 %
		LRT ($\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2$)	**99.5 %**	0 %	0.5 %	0 %
		LRT ($\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$)	**99.5 %**	0 %	0.5 %	0 %
(ii)	**GIM**	AIC	**16.0 %**	0 %	84.0 %	0 %
		LRT (χ^2_2)	**6.5 %**	0 %	93.5 %	0 %
		LRT ($\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2$)	**11.5 %**	0 %	88.5 %	0 %
		LRT ($\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$)	**15.5 %**	0 %	84.5 %	0 %
(iii)	**IIM**	AIC	2.0 %	**98.0 %**	0 %	0 %
		LRT (χ^2_2)	2.0 %	**98.0 %**	0 %	0 %
		LRT ($\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2$)	2.0 %	**98.0 %**	0 %	0 %
		LRT ($\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$)	2.0 %	**98.0 %**	0 %	0 %
(iv)	**secondary contact**	AIC	4.0 %	0 %	**96.0 %**	0 %
		LRT (χ^2_2)	2.0 %	0 %	**98.0 %**	0 %
		LRT ($\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2$)	2.0 %	0 %	**98.0 %**	0 %
		LRT ($\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$)	3.5 %	0 %	**96.5 %**	0 %
(v)	**isolation**	AIC	0 %	2.0 %	2.0 %	**96.0 %**
		LRT (χ^2_2)	0 %	0 %	0.5 %	**99.5 %**
		LRT ($\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2$)	0 %	0 %	0.5 %	**99.5 %**
		LRT ($\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$)	0 %	1.0 %	0.5 %	**98.5 %**

Results of the different model selection procedures for 200 data sets simulated under each of scenarios (i) to (v). The four models shown in Figure 1 were fitted to each simulated data set. For each scenario, the percentages shown are the proportions of data sets for which each of the four models was selected as the best-fitting model; the proportion of data sets for which the correct model was selected is shown in bold. For each simulated data set, model selection was performed using the methods described in Subsection 2.4. The method 'AIC' consists of selecting the model with the best AIC score. The method 'LRT' consists of a sequence of Likelihood Ratio Tests at an overall significance level of 5%, using the distribution shown in parentheses as the null distribution.

data sets simulated from the isolation model (scenario (v)); similarly, for data simulated from the IIM model (scenario (iii)) or the model of secondary contact (scenario (iv)), a type 1 error was made (falsely rejecting the true model in favour of the GIM model) in only 2% or 3.5% of cases, respectively – less than the 5% type 1 error rate that would be expected if we were able to use the precise null distribution of the test statistic. Compared to the other two distributions considered, the use of $\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$ gives somewhat higher power to detect the small level of gene flow in the intermediate stage of the model in scenario (ii).

4 Discussion

In this paper we have presented a Maximum-Likelihood method to estimate the parameters of a 'generalised isolation-with-migration model' from data consisting of the numbers of nucleotide differences between one pair of DNA sequences at each of a large number of independent loci. Our method is computationally very fast and can also be used to easily compare the fit of the four models depicted in Figure 1, thus making it possible to distinguish between these different evolutionary scenarios. In Janko et al. (2018) we applied our method (slightly simplified, with symmetric migration rates) to data from different species of *Cobitis* (spined loaches) to try to reconstruct the evolutionary history of these species. The results suggested that extensive historical gene flow occurred between *C. elongatoides* and the common ancestor of *C. taenia*, *C. tanaitica* and *C. pontica*, and that this was followed by reproductive isolation of all these species. In the previous sections we have presented the mathematical derivations underlying our method, together with a simulation study to evaluate the performance of our method under a number of different evolutionary scenarios with decreasing or increasing gene flow, or in the absence of gene flow.

For the vast majority of data sets simulated from scenarios (i), (iii), (iv) and (v) in Section 3, our method correctly identified the 'true' model. However, for data simulated from a scenario of increasing gene flow (scenario (ii)), there was little power to detect the very small amount of gene flow that had occurred in the intermediate stage of the model. To further investigate whether the lack of power in this particular scenario is due to the subsequent increase in the level of gene flow, or whether there is more generally a lack of power to detect a small amount of gene flow that occurred a long time ago,

we simulated 200 data sets from a model with the same parameters as in scenario (ii) except for the contemporary migration rates, which were set to 0:

- scenario (vi): $\theta_0 = 3$, $\theta_1 = 2$, $\theta_2 = 4$, $\theta'_1 = 3$, $\theta'_2 = 6$, $T_1 = 4$, $V = 4$, with migration parameters $M_1{}^* = 0.04$ and $M_2{}^* = 0.08$ decreasing to $M'_1{}^* = M'_2{}^* = 0$;

the number of loci in each simulated data set, and the relative mutation rates of the different loci, were also as in Section 3. We fitted the four models shown in Figure 1 to each of the simulated data sets and applied the model selection procedures described in Subsection 2.4. The results in Table 3 show that, in contrast with scenario (ii), the power to detect the small amount of historical gene flow in this new scenario (vi) was high: our method returned the correct IIM model for 92% of the simulated data sets

Table 3:
Model selection for simulated data: Results

simulation scenario	**true model**	**method**	**best-fitting model**			
			GIM	**IIM**	**secondary contact**	**isolation**
(vi)	**IIM**	AIC	0 %	**92.0 %**	2.5 %	5.5 %
as in scenario (ii)		LRT (χ^2_2)	0 %	**79.0 %**	2.0 %	19.0 %
but with		LRT ($\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2$)	0 %	**83.5 %**	2.0 %	14.5 %
$M'_1{}^* = M'_2{}^* = 0$		LRT ($\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$)	0 %	**88.0 %**	2.5 %	9.5 %
(vii)	**GIM**	AIC	**29.0 %**	0 %	71.0 %	0 %
as in scenario (ii)		LRT (χ^2_2)	**15.5 %**	0 %	84.5 %	0 %
but with		LRT ($\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2$)	**21.0 %**	0 %	79.0 %	0 %
$M_1{}^* = 0.08$ *and* $M_2{}^* = 0.16$		LRT ($\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$)	**28.0 %**	0 %	72.0 %	0 %
(viii)	**GIM**	AIC	**68.0 %**	0 %	32.0 %	0 %
as in scenario (ii)		LRT (χ^2_2)	**55.0 %**	0 %	45.0 %	0 %
but with		LRT ($\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2$)	**60.5 %**	0 %	39.5 %	0 %
$V = 8$		LRT ($\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$)	**65.5 %**	0 %	34.5 %	0 %

Results of the different model selection procedures for 200 data sets simulated under each of scenarios (vi) to (viii). The four models shown in Figure 1 were fitted to each simulated data set. For each scenario, the percentages shown are the proportions of data sets for which each of the four models was selected as the best-fitting model; the proportion of data sets for which the correct model was selected is shown in bold. For each simulated data set, model selection was performed using the methods described in Subsection 2.4. The method 'AIC' consists of selecting the model with the best AIC score. The method 'LRT' consists of a sequence of Likelihood Ratio Tests at an overall significance level of 5%, using the distribution shown in parentheses as the null distribution.

when the best-fitting model was selected by means of AIC scores, and for 83.5% or 88% of the simulated data sets when a sequence of Likelihood Ratio tests was used with, respectively, the mixtures $\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2$ or $\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$ instead of the precise null distribution. Thus it appears that the lack of power to detect the small amount of historical gene flow in scenario (ii) was specifically due to the subsequent increase in the level of gene flow: the larger level of contemporary gene flow appears to mask the signal from the small amount of earlier gene flow. In order to assess to what extent the power of our method in this type of scenario might improve if the level of historical gene flow was larger, or if this low level of historical gene flow lasted for a longer period of time, we also simulated 200 data sets from each of the following two modifications of scenario (ii):

- doubling the amount of historical gene flow:
  scenario (vii): $\theta_0 = 3$, $\theta_1 = 2$, $\theta_2 = 4$, $\theta_1' = 3$, $\theta_2' = 6$, $T_1 = 4$, $V = 4$, with $M_1^* = 0.08$ and $M_2^* = 0.16$ increasing to $M_1'^* = 0.2$ and $M_2'^* = 0.4$;

- doubling the duration of the intermediate time period:
  scenario (viii): $\theta_0 = 3$, $\theta_1 = 2$, $\theta_2 = 4$, $\theta_1' = 3$, $\theta_2' = 6$, $T_1 = 4$, $V = 8$, with $M_1^* = 0.04$ and $M_2^* = 0.08$ increasing to $M_1'^* = 0.2$ and $M_2'^* = 0.4$.

We found that while doubling the rate of historical gene flow only led to a modest improvement in power, doubling the duration of the intermediate time period resulted in considerably higher power to detect this historical gene flow (see Table 3).

For each of the Likelihood Ratio tests in the model selection procedure set out in Subsection 2.4, the asymptotic null distribution of the LRT statistic is a mixture of χ^2_ν distributions ($\nu = 0, 1, 2$), but the precise coefficients of χ^2_0 and χ^2_2 in the mixture are not easy to compute; the coefficient of χ^2_1 in the mixture is $\frac{1}{2}$. Whilst the use of χ^2_2 or $\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2$ instead of the precise asymptotic null distribution is obviously conservative, the simulation results in Section 3 suggest that the use of $\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$ also appears to be conservative, and results in somewhat higher power compared to the other two distributions. QQ-plots comparing the quantiles of the null distribution of the LRT statistic Λ obtained for simulated data with the theoretical quantiles of $\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$, for each of the four Likelihood Ratio tests in our model selection procedure, confirm that the use of $\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$ instead of the correct null distribution in these tests is indeed conservative - these plots are shown in the Appendix. Nevertheless, further work is needed to establish whether the use of $\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$ is always conservative

when testing whether two migration rates are both 0 in a GIM model or in one of its variants, whatever the true parameter values. Further work to more easily compute the precise asymptotic null distribution of the LRT statistic would also be useful, as it would enhance the power of the model selection procedure, but this is beyond the scope of the present study.

The work presented in this paper assumes the infinite sites model of mutation. While it would be straightforward to adapt our method to the Jukes-Cantor model of mutation (Jukes and Cantor, 1969) by combining our result (11) for the GIM model with equation (3) of Lohse et al. (2011), such implementation is left for future work. Extensions to more complex mutation models should also be possible but would require more effort.

Our ML method assumes that there is no recombination within loci and free recombination between loci. These assumptions, although commonly made in the literature, may often be violated for real data sets. In the 'Discussion' Section of Costa and Wilkinson-Herbots (2017) we examined in detail the effects of such potential violations of our assumptions on the accuracy of ML estimates obtained for the IIM model; parameter estimates obtained for the GIM model can be expected to be affected in a similar way. In particular, while non-negligible recombination within loci may lead to biased estimates, linkage between loci should not cause bias but – unless properly accounted for – would lead to underestimation of the uncertainty surrounding the estimates obtained (see also Baird, 2015; Lohse et al., 2016). We refer to Costa and Wilkinson-Herbots (2017) for an illustration of how linkage disequilibrium, and model misspecification more generally, can be accounted for in practice. In particular, we illustrated there how for a data set made up of clusters of loci, where loci within clusters are subject to linkage disequilibrium but linkage between clusters is negligible, robust standard errors and confidence intervals can readily be obtained by using an estimate of the inverse of the Godambe information matrix (the robust sandwich estimator of the parameter covariance matrix) rather than the inverse of the Fisher information matrix (Chandler and Bate, 2007; Varin, 2008; Jesus and Chandler, 2011), while robust likelihood ratio tests can be performed by using the scaled and shifted χ^2 distribution given in equation (3.6) of Jesus and Chandler (2011) as the null distribution. Alternatively, recombination and linkage disequilibrium can be accounted for by means of a parametric bootstrap (for example, Lohse et al., 2016), but this approach is computationally intensive and the results will inevitably depend on the

recombination rate assumed, and on any other assumptions made such as homogeneity of the recombination rate along the genome.

While our ML method accommodates mutation rate heterogeneity between loci (whether due to variation in sequence length or variation in the mutation rate per site, or both), it assumes that accurate estimates of the relative mutation rates of the different loci are available and treats these rates as known scalars. In practice, these scalars are usually estimated using outgroup sequences, and our method does not currently account for uncertainty about the relative mutation rates. Wang and Hey (2010) conducted a simulation study investigating the accuracy of ML estimates of the parameters of the IM model (where, as is the case in our method, the parameter estimates were obtained treating the relative mutation rates as known scalars), when the relative mutation rates of the different loci were in fact estimated using the distance between one DNA sequence from each of two outgroups (their 'fixed-rate' method). They also proposed an 'all-rate' method whereby the divergence between these two outgroup sequences at each locus is considered part of the data, and the relative mutation rates at the different loci are assigned a probability distribution (either gamma or uniform in their simulations) which is integrated over. Whereas our current implementation of the GIM model corresponds to the 'fixed-rate' method, it should be feasible to extend our method to incorporate the 'all-rate' method and to account for uncertainty about the relative mutation rates in that way.

Our method is suitable for DNA sequences of intermediate length, typically of the order of hundreds of base pairs. For very short sequences ($<<$ 100 bp), the infinite sites model may provide a poor approximation, whereas for very long sequences ($>>$ 1000 bp), intra-locus recombination cannot be neglected.

For mathematical simplicity and computational speed, our method uses data consisting of the number of nucleotide differences between one pair of DNA sequences at each of a large number of independent loci. It would be of interest to extend our method to accommodate somewhat larger samples of sequences at each locus, as this would be more natural and less wasteful of data in practice. The easiest way to include larger samples of sequences at each locus might be to use all pairwise subsamples and maximise the resulting composite marginal loglikelihood. However, because different pairs of sequences from the same locus are not independent, standard errors obtained with our code (based

on the Fisher information matrix) would in that case underestimate the true amount of uncertainty about the parameter estimates, and standard asymptotic results regarding the null distribution of the Likelihood Ratio test statistic would no longer apply. Instead, standard errors should be based on an estimate of the Godambe information matrix, and model selection criteria should be adjusted to apply to composite marginal likelihoods (see, for example, Varin et al., 2011, for a review); alternatively, simulation-based methods can be used to quantify uncertainty and perform model selection. It should also be possible to extend our derivation of the likelihood to somewhat larger numbers of sequences per locus, either by using an approach similar to that of Andersen et al. (2014) for sequences from two diploid individuals (one from each population), or by using an approach similar to that of Kumagai and Uyenoyama (2015) to derive the likelihood of a data set where the observation at each locus consists of the number of segregating sites in a sample of more than two sequences.

Having derived an explicit expression for the likelihood of a data set consisting of the numbers of nucleotide differences between one pair of DNA sequences at each of a large number of independent loci, it was natural to use ML methods for parameter estimation and model selection rather than a Bayesian approach, as maximising this likelihood is straightforward and requires very little computing time. If desired, it should nevertheless be possible to use our results for the likelihood as a building block to develop an analogous Bayesian method. Yang and Zhu (2018) pointed out some fundamental problems arising in Bayesian model selection when all models considered are misspecified, which is typically the case in evolutionary genetics. An additional advantage of using a ML framework may therefore be that the effects of model misspecification in this context are somewhat better understood and less difficult to account for.

Acknowledgements

We thank Karel Janko for valuable discussions which motivated the work presented in this paper. We also thank the anonymous reviewers for their constructive comments and suggestions. This work was supported by the UK Engineering and Physical Sciences Research Council [grant number EP/K502959/1].

Declarations of interest

None.

Appendix

For each of the Likelihood Ratio tests in the model selection procedure set out in Subsection 2.4, the asymptotic null distribution of the LRT statistic is a mixture of χ^2_ν distributions ($\nu = 0, 1, 2$), but the precise coefficients of χ^2_0 and χ^2_2 in the mixture are not easy to compute; the coefficient of χ^2_1 in the mixture is $\frac{1}{2}$. The simulation results in Section 3 suggest that the use of $\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$ instead of the correct null distribution appears to be conservative. In this Appendix we further examine whether that is indeed the case. To this end, QQ-plots were constructed (see Figure 4) comparing the quantiles of the null distribution of the LRT statistic Λ obtained for simulated data with the theoretical quantiles of $\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$, for each of the four Likelihood Ratio tests in our model selection procedure:

(a) the isolation model versus the IIM model;

(b) the isolation model versus the model of secondary contact;

(c) the IIM model versus the GIM model;

(d) the model of secondary contact versus the GIM model.

The QQ-plots are based on the 200 data sets that were simulated in Section 3 under the null hypothesis in each case: the isolation model (scenario (v)) for plots (a) and (b), the IIM model (scenario (iii)) for plot (c), and the model of secondary contact (scenario (iv)) for plot (d); full details of the simulations, and the parameter values used, are given in Section 3.

For the tests in (a), (b) and (c), the QQ-plots in Figure 4 confirm that the use of $\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$ instead of the correct null distribution is conservative, as in each of these plots all points lie below the diagonal red line, i.e. the quantiles of the (simulated) null distribution of the LRT statistic Λ are smaller than the corresponding quantiles of $\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$. In the QQ-plot in (d), the two most extreme points in the top right corner lie somewhat above the diagonal red line, and therefore additional simulations were carried out to assess whether this indicates non-conservativeness or whether this is merely due to chance: a further 300 data sets were simulated from scenario (iv) (the model of secondary contact), in addition to the 200 data sets already simulated from this scenario in Section 3. We fitted both the model of secondary contact and the full GIM

model to each simulated data set and computed the value of the LRT statistic Λ for the test of the model of secondary contact (H_0) against the GIM model (H_1). Figure 5 shows the QQ-plot of the distribution of Λ against the mixture $\frac{1}{4}\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{1}{4}\chi_2^2$, where the quantiles of Λ were computed using all 500 simulated observations of the LRT statistic. This plot indicates that, for the Likelihood Ratio test in (d), the use of $\frac{1}{4}\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{1}{4}\chi_2^2$ instead of the precise null distribution of Λ is also conservative.

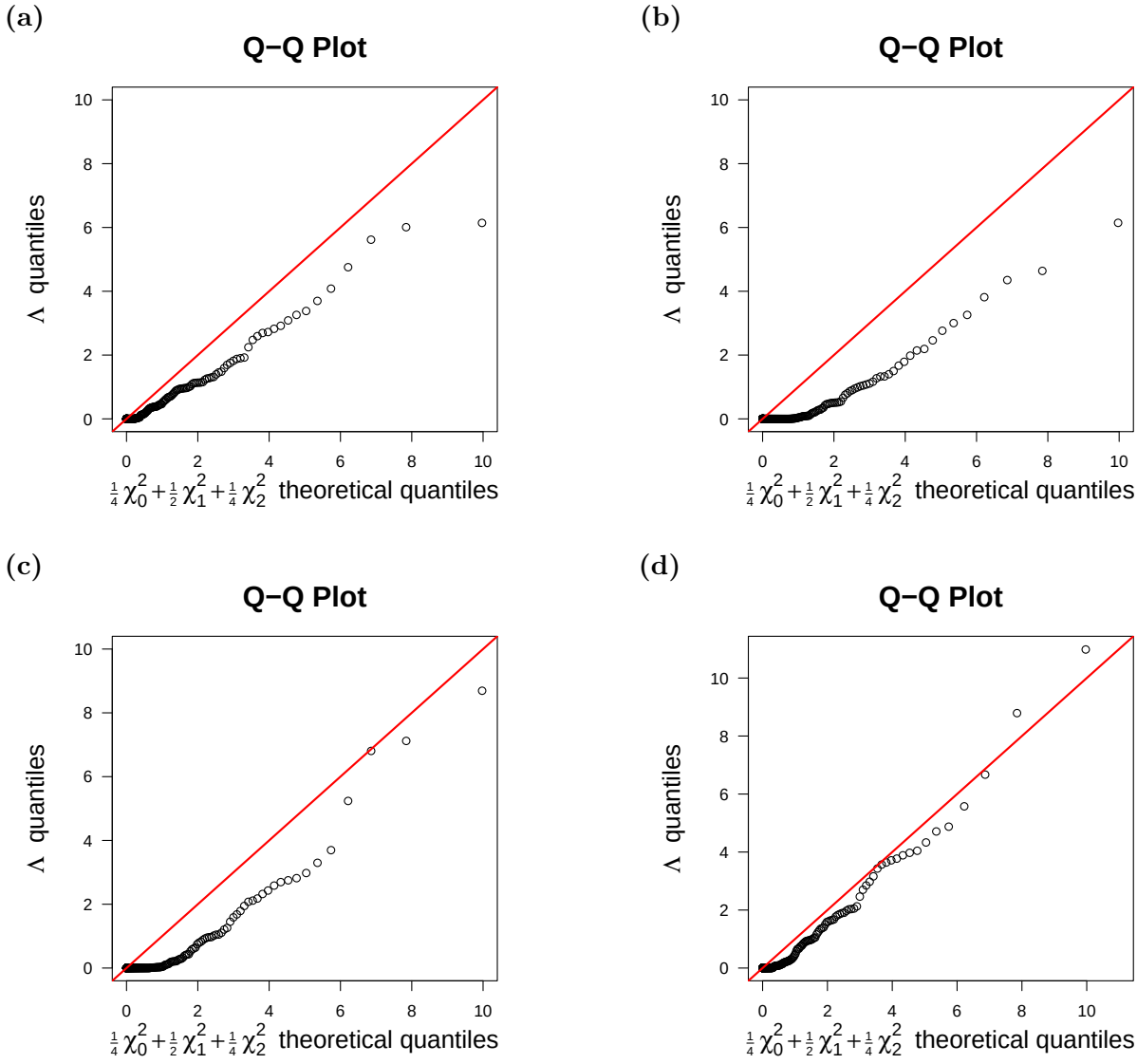

Figure 4: QQ-plots of the null distribution of the LRT statistic Λ (obtained for simulated data) against the mixture $\frac{1}{4}\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{1}{4}\chi_2^2$. Plots (a) and (b) are, respectively, for the LRT of the isolation model (H_0) against the IIM model (H_1), and the LRT of the isolation model (H_0) against the model of secondary contact (H_1), for 200 data sets simulated from an isolation model (scenario (v) in Section 3). Plot (c) is for the LRT of the IIM model (H_0) against the GIM model (H_1), for 200 data sets simulated from an IIM model (scenario (iii) in Section 3). Plot (d) is for the LRT of the model of secondary contact (H_0) against the GIM model (H_1), for 200 data sets simulated from a model of secondary contact (scenario (iv) in Section 3). The line $y = x$ is also shown (in red) for ease of comparison.

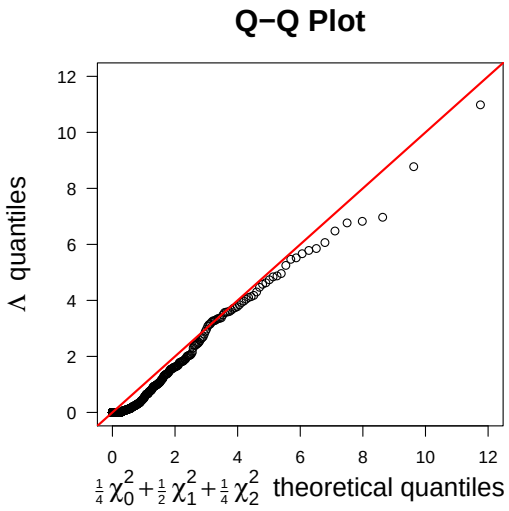

Figure 5: QQ-plot of the null distribution of the LRT statistic Λ (obtained for simulated data) against the mixture $\frac{1}{4}\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{1}{4}\chi_2^2$, for the LRT of the model of secondary contact (H_0) against the GIM model (H_1). The quantiles of Λ were based on 500 data sets simulated from a model of secondary contact (scenario (iv) in Section 3). The line $y = x$ is also shown (in red) for ease of comparison.

Bibliography

Akaike, H. (1972). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *Proc. 2nd Int. Symp. Information Theory*, pp. 267–281.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control AC-19*, 716–723.

Andersen, L., T. Mailund, and A. Hobolth (2014). Efficient computation in the IM model. *Journal of Mathematical Biology 68*(6), 1423–1451.

Baird, S. J. E. (2015). Exploring linkage disequilibrium. *Molecular Ecology Resources 15*, 1017–1019.

Becquet, C. and M. Przeworski (2007). A new approach to estimate parameters of speciation models with application to apes. *Genome Research 17*, 1505–1519.

Beeravolu, C. R., M. J. Hickerson, L. A. F. Frantz, and K. Lohse (2018). Able: blockwise site frequency spectra for inferring complex population histories and recombination. *Genome Biology 19*, 145.

Chandler, R. E. and S. Bate (2007). Inference for clustered data using the independence loglikelihood. *Biometrika 94*, 167–183.

Chen, H. (2012). The joint allele frequency spectrum of multiple populations: A coalescent theory approach. *Theoretical Population Biology 81*, 179–195.

Costa, R. J. and H. Wilkinson-Herbots (2017). Inference of gene flow in the process of speciation: An efficient maximum-likelihood method for the isolation-with-initial-migration model. *Genetics 205*, 1597–1618.

Dalquen, D. A., T. Zhu, and Z. Yang (2017). Maximum likelihood implementation of an isolation-with-migration model for three species. *Systematic Biology 66*, 379–398.

Excoffier, L., I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll (2013). Robust demographic inference from genomic and snp data. *PLoS Genetics 9*, e1003905.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection* (1st ed.). Oxford: Clarendon Press.

Flouri, T., X. Jiao, B. Rannala, and Z. Yang (2020). A bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Molecular Biology and Evolution 37*, 1211–1223.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante (2009). Inferring the joint demographic history of multiple populations from multidimentional snp frequency data. *PLoS Genetics 5*, e1000695.

Hearn, J., G. N. Stone, L. Bunnefeld, J. A. Nicholls, N. H. Barton, and K. Lohse (2014). Likelihood-based inference of population history from low-coverage de novo genome assemblies. *Molecular Ecology 23*, 198–211.

Herbots, H. M. (1997). The structured coalescent. In P. Donnelly and S. Tavar/'e (Eds.), *Progress in population genetics and human evolution*, Volume 87 of *IMA Volumes in Mathematics and its Applications*, pp. 231–255. Springer-Verlag.

Hey, J. (2005). On the Number of New World Founders: A Population Genetic Portrait of the Peopling of the Americas. *PLoS Biol 3*, e193.

Hey, J. (2010). Isolation with Migration Models for More Than Two Populations. *Molecular Biology and Evolution 27*, 905–920.

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

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

Innan, H. and H. Watanabe (2006). The effect of gene flow on the coalescent time in the human-chimpanzee ancestral population. *Molecular Biology and Evolution 23*, 1040–1047.

Janko, K., J. Pačes, H. Wilkinson-Herbots, R. J. Costa, J. Roslein, P. Drozd, N. Iakovenko, J. Rídl, M. Hroudová, J. Kočí, R. Reifová, V. Šlechtová, and L. Choleva

(2018). Hybrid asexuality as a primary postzygotic barrier between nascent species: On the interconnection between asexuality, hybridization and speciation. *Molecular Ecology 27*, 248–263.

Jesus, J. and R. E. Chandler (2011). Estimating functions and the generalized method of moments. *Interface focus 1*, 871–885.

Jukes, T. H. and C. R. Cantor (1969). Evolution of protein molecules. In H. N. Munro (Ed.), *Mammalian Protein Metabolism*, pp. 21–123. New York: Academic Press.

Keightley, P. D., R. W. Ness, D. L. Halligan, and P. R. Haddrill (2014). Estimation of the spontaneous mutation rate per nucleotide site in a *drosophila melanogaster* full-sib family. *Genetics 196*, 313–320.

Kern, A. D. and J. Hey (2017). Exact calculation of the joint allele frequency spectrum for isolation with migration models. *Genetics 207*, 241–253.

Kingman, J. F. C. (1982a). Exchangeability and the evolution of large populations. In G. Koch and F. Spizzichino (Eds.), *Exchangeability in Probability and Statistics*, Proceedings of the International Conference on Exchangeability in Probability and Statistics, pp. 97–112. North-Holland Elsevier.

Kingman, J. F. C. (1982b). On the Genealogy of Large Populations. *Journal of Applied Probability 19*, 27–43.

Kingman, J. F. C. (1982c). The Coalescent. *Stochastic Processes and Their Applications 13*(3), 235–248.

Kozakai, R., A. Shimizu, and M. Notohara (2016, 06). Convergence to the structured coalescent process. *J. Appl. Probab. 53*, 502–517.

Kumagai, S. and M. K. Uyenoyama (2015). Genealogical histories in structured populations. *Theoretical Population Biology 102*, 3–15.

Lohse, K., M. Chmelik, S. H. Martin, and N. H. Barton (2016). Efficient strategies for calculating blockwise likelihoods under the coalescent. *Genetics 202*, 775–786.

Lohse, K. and L. A. F. Frantz (2014). Neandertal Admixture in Eurasia Confirmed by Maximum-Likelihood Analysis of Three Genomes. *Genetics 196*, 1241–1251.

Lohse, K., R. J. Harrison, and N. H. Barton (2011). A general method for calculating likelihoods under the coalescent process. *Genetics 189*, 977–987.

Lohse, K., B. Sharanowski, and G. N. Stone (2010). Quantifying the pleistocene history of the oak gall parasitoid *Cecidostiba fungosa* using twenty intron loci. *Evolution 64*, 2664–2681.

Lukić, S. and J. Hey (2012). Demographic inference using spectral methods on snp data, with an analysis of the human out-of-africa expansion. *Genetics 192*, 619–639.

Maddison, W. P. and L. L. Knowles (2006). Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology 55*, 21–30.

Mailund, T., A. E. Halager, M. Westergaard, J. Y. Dutheil, K. Munch, L. N. Andersen, G. Lunter, K. Prüfer, A. Scally, A. Hobolth, and M. H. Schierup (2012). A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. *PLoS Genetics 8*, e1003125.

Naduvilezhath, L., L. E. Rose, and D. Metzler (2011). Jaatha: a fast composite-likelihood approach to estimate demographic parameters. *Molecular Ecology 20*, 2709–2723.

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

Notohara, M. (1990). The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology 29*(1), 59–75.

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

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Roux, C., C. Fraïsse, J. Romiguier, Y. Anciaux, N. Galtier, and N. Bierne (2016). Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS Biol 14*, e2000234.

Self, S. G. and K.-Y. Liang (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association 82*, 605–610.

Silvapulle, M. J. and P. K. Sen (2005). *Constrained Statistical Inference: Order, Inequality, and Shape Constraints*. John Wiley & Sons.

Sousa, V. C. and J. Hey (2013). Understanding the origin of species with genome-scale data: modelling gene flow. *Nat Rev Genet 14*, 404–414.

Takahata, N. (1988). The coalescent in two partially isolated diffusion populations. *Genetical Research 52*(3), 213–222.

Takahata, N., Y. Satta, and J. Klein (1995). Divergence time and population size in the lineage leading to modern humans. *Theoretical Population Biology 48*, 198 – 221.

Terhorst, J. and Y. S. Song (2015). Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proceedings of the National Academy of Sciences of the United States of America 112*, 7677–7682.

Varin, C. (2008). On composite marginal likelihoods. *AStA Adv. Stat. Anal. 92*, 1–28.

Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. *Statistica Sinica 21*, 5–42.

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

Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology 7*, 256–276.

Wilkinson-Herbots, H. (2012). The distribution of the coalescence time and the number of pairwise nucleotide differences in a model of population divergence or speciation with an initial period of gene flow. *Theoretical Population Biology 82*, 92–108.

Wilkinson-Herbots, H. (2015). A fast method to estimate speciation parameters in a model of isolation with an initial period of gene flow and to test alternative evolutionary scenarios. *ArXiv e-prints*. URL: *http://arxiv.org/abs/1511.05478* .

Wilkinson-Herbots, H. M. (2008). The distribution of the coalescence time and the number of pairwise nucleotide differences in the isolation with migration model. *Theoretical Population Biology 73*, 277 – 288.

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

Yang, Z. (2002). Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics 162*, 1811–1823.

Yang, Z. and T. Zhu (2018). Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees. *PNAS 115*, 1854–1859.

Zhu, T. and Z. Yang (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.