Coalescence times for three genes provide sufficient information to distinguish population structure from population size changes

Simona Grusea, Willy Rodríguez, Didier Pinchon, Lounès Chikhi, Simon Boitard, Olivier Mazet

Coalescence times for three genes provide sufficient information to distinguish population structure from population size changes

Simona Grusea*,[a], Willy Rodríguez[a], Didier Pinchon[b], Lounès Chikhi[c,d,e], Simon Boitard[f], and Olivier Mazet[a]

[a]Institut de Mathématiques de Toulouse, Université de Toulouse, Institut National des Sciences Appliquées, F-31077 Toulouse, France
[b]Institut de Mathématiques de Toulouse, Université de Toulouse, F-31077 Toulouse, France
[c]CNRS, Université Paul Sabatier, ENFA, UMR 5174 EDB (Laboratoire Évolution & Diversité Biologique), Bât. 4R1, F-31062 Toulouse, France
[d]Université de Toulouse, UPS, EDB, F-31062 Toulouse, France
[e]Instituto Gulbenkian de Ciência, Rua da Quinta Grande, No. 6, P-2780-156 Oeiras, Portugal
[f]GenPhySE, Université de Toulouse, INRA, INPT, INP-ENVT, Castanet Tolosan, France

November 9, 2017

*Corresponding author. Email: grusea@insa-toulouse.fr

Abstract

The increasing amount of genomic data currently available is expanding the horizons of population genetics inference. A wide range of methods have been published allowing to detect and date major changes in population size during the history of species. At the same time, there has been an increasing recognition that population structure can generate genetic data similar to those generated under models of population size change. Recently, Mazet et al. [2016] introduced the idea that, for any model of population structure, it is always possible to find a panmictic model with a particular function of population size change having an identical distribution of T_2 (the time of the first coalescence for a sample of size two). This implies that there is an identifiability problem between a panmictic and a structured model when we base our analysis only on T_2. A natural question that deserves to be explored is whether and when this identifiability problem disappears for larger sample sizes. In this paper, based on an analytical study of the rate matrix (or Q-matrix) of the ancestral lineage process, we obtain new theoretical results about the joint distribution of the coalescence times (T_3, T_2) for a sample of three haploid genes in a n-island model with constant size. In particular, we show that this distribution is always different from the analogous one obtained in a panmictic population, for any scenario of population size-change. Even if, for any $k \geq 2$, it is always possible to find a size-change scenario for a panmictic population such that the marginal distribution of T_2 is exactly the same as in a n-island model with constant population size, we show that the joint distribution of the coalescence times (T_3, T_2) for a sample of three genes contains enough information to distinguish between a panmictic population and a n-island model of constant size.

Keywords: IICR (inverse instantaneous coalescence rate); population structure; population size change; demographic history; rate matrix; structured coalescent.

1 Introduction

Coalescent theory was developed in the 1980s by theoretical population geneticists interested in the statistical properties of gene trees [Kingman, 1982, Hudson, 1983, Tajima, 1983]. By focusing on the properties of *samples* within populations, coalescent theory allowed a significant reduction in the computational cost of simulations [Hudson et al., 1990, Hudson, 2002]. Also, by stressing the importance of backward inference, it allowed new insights in our understanding of the shapes of gene genealogies obtained from real species [Slatkin, 1991, Rogers and Harpending, 1992]. This change in focus (a backward sample view instead of a forward population view) allowed the development of new methods to infer the demographic history of populations and species [Beaumont, 1999, Hudson, 2002, Nielsen and Wakeley, 2001] and the detection of ancient population size changes in many species [Storz and Beaumont, 2002, Goossens et al., 2006, Quéméré et al., 2012]. However, there has been an increasing recognition that population structure can generate false signatures of population size change.

Indeed, coalescent theory predicts that gene trees obtained from bottlenecked populations may be similar in shape to those obtained from structured populations [Wakeley, 1999, Storz and Beaumont, 2002, Mazet et al., 2015]. Significant work has been done to find a solution to this inference problem, most of which is based on simulations rather than analytical work [Beaumont, 2004, Chikhi et al., 2010, Peter et al., 2010, Heller et al., 2013]. One reason for this is that the coalescent theory of structured population is difficult. Indeed, while it is straightforward to simulate coalescent times for nearly any model of population structure and any sampling scheme [Hudson, 2002], few theoretical results exist regarding the distributions of these times for sample sizes above two. Thus, while there has been significant progress in our understanding of coalescent theory for a wide variety of models, several difficult problems in demographic inference could still be addressed using analytical approaches. For instance, Mazet et al. [2015] used the distribution of the coalescence time T_2 for a sample of two haploid genomes (or one diploid individual) to separate two models for which these distributions could be derived. The first model was a structured model, the n-island model of Wright [1931], whereas the second model was a simple panmictic model with only one stepwise population size change. In that particular case they showed that the distributions were different and could thus be separated with a reasonably limited number of independent values of T_2.

This provided a promising result, since it suggested that genomic data from a single diploid individual could be enough to separate one model of population structure from a model of population size change. Such a method would contribute to solve problems pointed in several simulation studies based on larger sample sizes [Chikhi et al., 2010, Peter et al., 2010, Heller et al., 2013]. However, in a more recent study, Mazet et al. [2016] have shown that, given the distribution of the coalescence time T_2 obtained under any model of population structure, there always exists a function $\lambda(\cdot)$ of population size-change which perfectly mimics this distribution. This function was derived for a sample of size two for the n-island model and called IICR, which stands for *inverse instantaneous coalescence rate*. In other words, the T_2 distribution alone does not allow distinguishing between a panmictic population whose size can vary arbitrarily and a structured population, however complex that structure may be, as they will have exactly the same IICR. Mazet et al. [2016] noted that this limitation might be overcome by increasing the sample size to more than two haploid genomes and including information from other coalescence times. They also noted that the sampling scheme could potentially also be used to separate two demographic models, because the IICR is a function of the sampling scheme, an issue that has been explored by Chikhi et al. [2018].

Altogether, obtaining the distribution of different T_k, for several sample sizes k and several sampling schemes, could in principle allow us determining whether these distributions differ significantly for models of population size changes and different models of population structure

Chikhi et al. [2018]. This would be crucial for demographic inference, in an era where genomic data are becoming increasingly available.

In this article we focus on the joint distribution of the coalescence times $(T_3^{(3)}, T_2^{(3)})$ for a sample of three haploid genes and show that it can be used to distinguish between panmixia with population size changes and a n-island model with constant size. Throughout this article, the superscript $^{(3)}$ identifies the fact that we consider a sample of 3 genes ; in particular, it allows distinguishing the second coalescence time $T_2^{(3)}$ for a sample of three genes from the coalescence time $T_2^{(2)}$ for a sample of only 2 genes, which we simply denote T_2.

In Section 2 we derive new theoretical results about the joint distribution of $(T_3^{(3)}, T_2^{(3)})$ in a n-island model with constant size, for $n \geq 3$. In particular, we explicitly diagonalise the rate matrix (also called Q-matrix) of the ancestral lineage process associated to this model and obtain closed analytic expressions for the transition probabilities. We also study in detail the distribution of $T_3^{(3)}$ and the population size change model mimicking this distribution, as previously done by Mazet et al. [2016] for a sample of size 2. In Section 3 we use the results obtained in Section 2 to compare the joint distributions of $(T_3^{(3)}, T_2^{(3)})$ for a panmictic population and a n-island model with $n \geq 3$, and demonstrate that these two distributions are always different, for any population size change and any sampling configuration.

Appendix A contains proofs of the results presented in Sections 2 and 3. In Appendix B we treat the special case of a 2-island model and give analogs for the main results presented in Sections 2 and 3.

2 Joint distribution of coalescence times for three genes

2.1 Panmictic model with population size changes

Consider a panmictic population whose population size history is represented by the function $\lambda(.)$, *i.e.* whereby the population size at time t in the past is given by $N(t) = N \cdot \lambda(t)$, with N being the present population size. We are interested in the coalescence times for a sample of three (haploid) genes. After rescaling time by units of N generations and taking $N \to \infty$, we let $T_3^{(3),\lambda}$ and $T_2^{(3),\lambda}$ denote the first and second coalescence times for the sample, under this model.

Using known results on the coalescent in populations of variable size (see for example Griffiths and Tavaré [1994]), the distribution of the first coalescence time, $T_3^{(3),\lambda}$, is given by

$$\mathbb{P}(T_3^{(3),\lambda} > t) = \exp\{-3\Lambda(t)\}, \tag{1}$$

where $\Lambda(t) = \int_0^t \frac{1}{\lambda(s)} ds$.

Conditional on $T_3^{(3),\lambda}$, the distribution of the second coalescence time $T_2^{(3),\lambda}$ is given by

$$\mathbb{P}(T_2^{(3),\lambda} > u | T_3^{(3),\lambda} = t) = \exp\{-(\Lambda(t+u) - \Lambda(t))\}.$$

The marginal distribution of $T_2^{(3),\lambda}$ is thus obtained by integrating over all possible values of t:

$$\begin{aligned}
\mathbb{P}(T_2^{(3),\lambda} > u) &= \int_0^\infty \mathbb{P}(T_2^{(3),\lambda} > u | T_3^{(3),\lambda} = t) f_{T_3^{(3),\lambda}}(t) dt \\
&= \int_0^\infty \frac{3}{\lambda(t)} \exp\{-3\Lambda(t)\} \exp\{-(\Lambda(t+u) - \Lambda(t))\} dt.
\end{aligned}$$

2.2 The symmetrical n-island model

Let us now consider the symmetrical n-island model, with $n \geq 3$ islands (also named subpopulations or demes), where N is the haploid size of each subpopulation, supposed constant over time. In the same way as in Wilkinson-Herbots [1998], we rescale time by units of N generations and take $N \to \infty$, in such a way that in this new continuous time scale, two genes in a single subpopulation have coalescence rate 1 (going backwards in time) and we call $M/2$ the scaled backward migration rate (i.e. the rate at which each gene leaves its subpopulation when we go backwards in time). In other words, if we call m the proportion of immigrant genes in each subpopulation (forward in time), we have $M = 2Nm$. For more details on the structured coalescent process see Herbots [1994], Wilkinson-Herbots [1998] and Notohara [1990].

As in the previous section, we are interested in the coalescence times for a sample of three genes. We denote $T_3^{(3),n,M}$ and $T_2^{(3),n,M}$ the first and second coalescence times for the sample, respectively. Every pair of genes in the same subpopulation may coalesce at rate 1 and every gene migrates at rate $M/2$.

Due to population structure, we now need to distinguish three different configurations for the sample (for the case where $n \geq 3$) : (1) the three genes were sampled from the same deme, (2) two genes were sampled from the same deme and the third gene was sampled from a different deme, or (3) the three genes were sampled from three different demes. For $i = 1, 2, 3$, we will denote $\mathbb{P}_i$ the conditional probability starting from sampling configuration i.

The special case of $n = 2$ islands needs a special treatment, since in this case the situation (3) cannot exist. We will treat this case in Appendix B.

2.2.1 The ancestral lineage process

Consider three (haploid) genes sampled in a population described by a symmetrical n-island model with $n \geq 3$ islands and migration parameter $M > 0$. In order to study the distribution of the first coalescence time $T_3^{(3),n,M}$ of these three genes, we introduce a Markovian jump process (called *ancestral lineage process*) describing the configuration in which the ancestral lineages of the three genes are at any time in the past, with the time going backwards, until their first coalescence event. The possible configurations, denoted by $i = 1, \ldots, 5$, are the following :

1. the three lineages are in the same island,
2. two lineages are in the same island and the third one is in a different island,
3. the three lineages are all in different islands,
4. there are only two ancestral lineages left and they are in the same island,
5. there are only two ancestral lineages left and they are in different islands.

The transition rate matrix Q of this Markovian jump process (also called Q-matrix, see Norris [1998] for definition) can be easily constructed (see Rodriguez [2016] and Rodriguez et al. [201X] for more details):

$$Q = \begin{pmatrix} -\frac{3M}{2} - 3 & \frac{3M}{2} & 0 & 3 & 0 \\ \frac{M}{2(n-1)} & -\frac{M(2n-3)}{2(n-1)} - 1 & \frac{M(n-2)}{n-1} & 0 & 1 \\ 0 & \frac{3M}{n-1} & -\frac{3M}{n-1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{2}$$

These transition rates are easily obtained using the fact that any two lineages among the three coalesce at rate 1 if they are in the same deme, and any of the three lineages migrates at rate

$\frac{M}{2}$; when a lineage migrates, it migrates to one of the other $n-1$ islands with equal probability. Given that we stop the process at the first coalescence event, the states 4 and 5 are absorbing.

Note that Herbots [1994] introduced the *structured coalescent* and gave an expression for the rate matrix Q in a general model of population structure. She considered an ancestral lineage process describing the number of ancestral lineages in each of the n-islands, backwards in time. In this article we only consider the case of a *symmetrical* n-island model, and thanks to the symmetries, we only need to consider the five configurations (or states) presented above and the *reduced* rate matrix Q given in Equation 2.

Let $P_t := e^{tQ}$, for $t > 0$, be the transition kernel of the above Markovian jump process (see again Norris [1998] for details). By definition, $P_t(i, j)$ represents the probability of the process being in state j at time t, conditional on starting in state i at time 0.

The cumulative distribution function (*cdf*) of the first coalescence time $T_3^{(3),n,M}$ can easily be expressed in terms of the matrix P_t, as follows :

$$\mathbb{P}_i(T_3^{(3),n,M} \leq t) = P_t(i, 4) + P_t(i, 5). \tag{3}$$

Moreover, using the fact that $\frac{d}{dt}P_t = P_tQ$, we have the relations

$$\begin{aligned}\frac{d}{dt}P_t(i, 4) &= 3P_t(i, 1),\\ \frac{d}{dt}P_t(i, 5) &= P_t(i, 2),\end{aligned}$$

from which we deduce the following expression for the density function of $T_3^{(3),n,M}$:

$$f_{T_3^{(3),n,M},i}(t) = \frac{d}{dt}\mathbb{P}_i(T_3^{(3),n,M} \leq t) = 3P_t(i, 1) + P_t(i, 2). \tag{4}$$

In order to obtain a closed form expression for P_t, and consequently for the distribution of $T_3^{(3),n,M}$ (see equation (3)), we study below the diagonalization of the rate matrix Q.

2.2.2 A closed form expression for P_t

The characteristic polynomial of Q is given by $\chi_Q(\mu) := \det(Q - \mu I_5) = -\mu^2 p(\mu)$, with

$$\begin{aligned}p(\mu) = \mu^3 &+ \frac{1}{2}\,\frac{8\,n - 8 + 5\,Mn}{n-1}\mu^2 + \frac{3}{2}\,\frac{Mn - 4\,M + 2\,n^2 - 4\,n + 3\,Mn^2 + 2 + M^2n^2}{(n-1)^2}\mu\\ &+ \frac{9}{2}\,\frac{M\,(2\,n + Mn - 2)}{(n-1)^2}.\end{aligned}$$

Thus, the matrix Q has the double eigenvalue 0. Concerning the other three eigenvalues, we can obtain the first lemma below:

Lemma 1. *The polynomial $p(\mu)$ has three distinct strictly negative real roots, $\mu_3 < \mu_2 < \mu_1 < 0$, such that*

$$\mu_3 < -2 - \frac{nM}{n-1} < \mu_2 < -\frac{3}{n} < \mu_1 < 0.$$

The proof of this lemma is given in Appendix A.

Using the Cardan-Viète method to express the roots of a third degree polynomial, such as exposed e.g. in Nickalls [1993], classical computations lead to the following explicit expressions for μ_1, μ_2 and μ_3:

Proposition 2. *Let us define the real numbers γ, K, x, r and α by*

$$\begin{aligned}
\gamma &= \frac{M}{n-1}, \\
K &= -\frac{5\gamma n + 8}{6}, \\
x &= (\gamma n - 2)(10\gamma^2 n^2 + 41\gamma n + 54\gamma + 40), \\
r &= \sqrt{7\gamma^2 n^2 + 26\gamma n - 72\gamma + 28}, \\
\alpha &= \frac{1}{3}\arccos(x/r^3).
\end{aligned}$$

Then the three negative real roots $\mu_3 < \mu_2 < \mu_1 < 0$ of $p(\mu)$ are given by

$$\mu_1 = \frac{r}{3}\cos(\alpha) + K, \quad \mu_2 = \frac{r}{3}\cos(\alpha + \frac{4\pi}{3}) + K, \quad \mu_3 = \frac{r}{3}\cos(\alpha + \frac{2\pi}{3}) + K.$$

Using the diagonalization of the rate matrix Q, we finally obtain the following expression for the transition kernel P_t :

Proposition 3. *The transition kernel P_t is given for any $t > 0$ by*

$$P_t = \sum_{i=1}^{3} e^{\mu_i t} A(\mu_i) + B, \tag{5}$$

where we denote

$$A(\mu) = \frac{1}{\delta(M,n,\mu)} \begin{bmatrix} \frac{3M^2u}{v} & 3(n-1)Mu & 3(n-1)(n-2)M^2 & \frac{9M^2u}{\mu v} & \frac{3(n-1)Mu}{\mu} \\ Mu & (n-1)uv & (n-1)(n-2)Mv & \frac{3Mu}{\mu} & \frac{(n-1)uv}{\mu} \\ 3M^2 & 3(n-1)Mv & \frac{3(n-1)(n-2)M^2v}{u} & \frac{9M^2}{\mu} & \frac{3(n-1)Mv}{\mu} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$B = \begin{bmatrix} 0 & 0 & 0 & b_1 & 1-b_1 \\ 0 & 0 & 0 & b_2 & 1-b_2 \\ 0 & 0 & 0 & b_2 & 1-b_2 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

with

$$\begin{aligned}
&\delta(M,n,\mu) = 2(n-1)^2 p'(\mu), \\
&u = u(M,n,\mu) = (n-1)\mu + 3M, \quad v = v(M,\mu) = 2\mu + 3(M+2), \\
&b_1 = b_1(M,n) = \frac{M + 2(n-1)}{nM + 2(n-1)}, \; b_2 = b_2(M,n) = \frac{M}{nM + 2(n-1)}.
\end{aligned}$$

The proof of the above proposition is deferred to Appendix A.

2.2.3 Joint distribution of $(T_3^{(3),n,M}, T_2^{(3),n,M})$

The joint distribution function of the two coalescence times for three genes sampled in configuration $i = 1, 2, 3$ can be obtained as follows:

$$\begin{aligned}\mathbb{P}_i(T_2^{(3),n,M} \leq u, T_3^{(3),n,M} \leq t) &= \mathbb{P}_i(T_2^{(3),n,M} \leq u, T_3^{(3),n,M} \leq t, C_s) \\ &\quad + \mathbb{P}_i(T_2^{(3),n,M} \leq u, T_3^{(3),n,M} \leq t, C_d),\end{aligned}$$

where C_s (respectively C_d) denotes the event that, after the first coalescence, the two remaining lineages are in the same deme (respectively in two different demes). Using the Markovian property of the ancestral lineage process and the fact that the states 4 and 5 are absorbing, we have

$$\begin{aligned}\mathbb{P}_i(T_2^{(3),n,M} \leq u, T_3^{(3),n,M} \leq t, C_s) &= F_{2,s}(u)P_t(i,4), \\ \mathbb{P}_i(T_2^{(3),n,M} \leq u, T_3^{(3),n,M} \leq t, C_d) &= F_{2,d}(u)P_t(i,5),\end{aligned}$$

where $F_{2,s}$ (resp. $F_{2,d}$) denotes (using the notations in Mazet et al. [2015]) the distribution function of the coalescence time $T_{2,s}$ (resp. $T_{2,d}$) for a sample of two genes taken in the same deme (resp. in two different demes).

We thus deduce

$$\mathbb{P}_i(T_2^{(3),n,M} \leq u, T_3^{(3),n,M} \leq t) = F_{2,s}(u)P_t(i,4) + F_{2,d}(u)P_t(i,5). \tag{6}$$

We can specify the above formula by plugging the expressions for $F_{2,s}$ and $F_{2,d}$ obtained in previous studies (see Herbots [1994] and Mazet et al. [2015]):

$$F_{2,s}(t) = \mathbb{P}(T_{2,s} \leq t) = \frac{a}{\alpha}\left(1 - e^{-\alpha t}\right) + \frac{1-a}{\beta}\left(1 - e^{-\beta t}\right), \tag{7}$$

$$F_{2,d}(t) = \mathbb{P}(T_{2,d} \leq t) = \frac{c}{\alpha}\left(1 - e^{-\alpha t}\right) - \frac{c}{\beta}\left(1 - e^{-\beta t}\right), \tag{8}$$

where

$$a = \frac{\gamma - \alpha}{\beta - \alpha}, \quad c = \frac{\gamma}{\beta - \alpha} \tag{9}$$

and $-\alpha$ and $-\beta$ are the roots of the polynomial $q(X) = X^2 + (1+n\gamma)X + \gamma$, whose discriminant equals $\Delta = (1+n\gamma)^2 - 4\gamma$, and therefore

$$\alpha = \frac{1}{2}\left(1 + n\gamma + \sqrt{\Delta}\right), \ \beta = \frac{1}{2}\left(1 + n\gamma - \sqrt{\Delta}\right), \tag{10}$$

with $\gamma = \frac{M}{n-1} = \alpha\beta$. Note that $-\alpha < -\beta < 0$ are the negative eigenvalues of the analogous rate matrix of the ancestral lineage process for a sample of size 2.

From Equation (6) we derive the conditional distribution of $T_2^{(3),n,M}$ given the value of $T_3^{(3),n,M}$, as follows

$$\begin{aligned}\mathbb{P}_i(T_2^{(3),n,M} \leq u | T_3^{(3),n,M} = t) &= \frac{F_{2,s}(u)\frac{d}{dt}P_t(i,4) + F_{2,d}(u)\frac{d}{dt}P_t(i,5)}{f_{T_3^{(3),n,M},i}(t)} \\ &= \frac{3F_{2,s}(u)P_t(i,1) + F_{2,d}(u)P_t(i,2)}{3P_t(i,1) + P_t(i,2)}.\end{aligned} \tag{11}$$

Moreover, by Equation (6), the marginal distribution function of $T_2^{(3),n,M}$ can be expressed as:

$$\mathbb{P}_i(T_2^{(3),n,M} \leq u) = F_{2,s}(u)\lim_{t\to\infty} P_t(i,4) + F_{2,d}(u)\lim_{t\to\infty} P_t(i,5).$$

Using the results in Proposition 3, we finally obtain explicitly the above limits and hence the distribution function of $T_2^{(3),n,M}$:

Proposition 4. *For every $u > 0$ we have:*

$$\mathbb{P}_1(T_2^{(3),n,M} \leq u) \;=\; \frac{M+2n-2}{Mn+2n-2}F_{2,s}(u) + \left(1 - \frac{M+2n-2}{Mn+2n-2}\right) F_{2,d}(u), \tag{12}$$

and for $i = 2$ and $i = 3$,

$$\mathbb{P}_i(T_2^{(3),n,M} \leq u) \;=\; \frac{M}{Mn+2n-2}F_{2,s}(u) + \left(1 - \frac{M}{Mn+2n-2}\right) F_{2,d}(u), \tag{13}$$

with $F_{2,s}$ and $F_{2,d}$ given by Equations (7) and (8).

The expected value of $T_2^{(3),n,M}$ can be deduced from the above proposition. Indeed, we have

$$\mathbb{E}_1(T_2^{(3),n,M}) = \frac{M+2n-2}{Mn+2n-2}\mathbb{E}(T_{2,s}) + \left(1 - \frac{M+2n-2}{Mn+2n-2}\right) \mathbb{E}(T_{2,d}).$$

Using the fact that $\mathbb{E}(T_{2,s}) = n$ and $\mathbb{E}(T_{2,d}) = n + \dfrac{n-1}{M}$ (see [Herbots, 1994]), we easily deduce that

$$\mathbb{E}_1(T_2^{(3),n,M}) = n + \frac{(n-1)^2}{Mn+2n-2}.$$

For $i = 2$ or $i = 3$, we get

$$\mathbb{E}_i(T_2^{(3),n,M}) = \frac{M}{Mn+2n-2}\mathbb{E}(T_{2,s}) + \left(1 - \frac{M}{Mn+2n-2}\right) \mathbb{E}(T_{2,d}),$$

and thus

$$\mathbb{E}_i(T_2^{(3),n,M}) = n + \frac{(n-1)^2(M+2)}{M(Mn+2n-2)},\ i = 2, 3.$$

Note that the same formulae were obtained in Herbots [1994] using Laplace transform methods.

2.2.4 Distribution of $T_3^{(3),n,M}$ and corresponding population size-change model

Equation (4) and Proposition 3 provide analytic expressions for the probability density function of $T_3^{(3),n,M}$ for each of the three sampling configurations. In guise of example, in Figure 1 we show the plot of the density of $T_3^{(3),n,M}$ for three genes sampled from the same deme (i.e. in sampling configuration $i = 1$), for $n = 10$ and $M = 1$ (in blue), and $M = 0.1$ (in red).

Considering a sample of two genes, Mazet et al. [2016] introduced a function $\lambda(\cdot)$ called the *inverse instantaneous coalescence rate* (IICR), which they defined as

$$\lambda(t) = \frac{\mathbb{P}(T_2 > t)}{f_{T_2}(t)}, \quad t \geq 0. \tag{14}$$

They showed that the distribution of their coalescence time T_2 can always be expressed as a function of $\lambda(\cdot)$, as follows:

$$\mathbb{P}(T_2 > t) = \exp\{-\Lambda(t)\}, \quad \Lambda(t) = \int_0^t \frac{1}{\lambda(s)} ds.$$

This expression holds for *any* model of population structure, but it is also exactly what we expect for a panmictic model whose population size along time is described by $N(t) = N(0)\lambda(t)$. This implies that for a structured model of any complexity, including thus the symmetrical

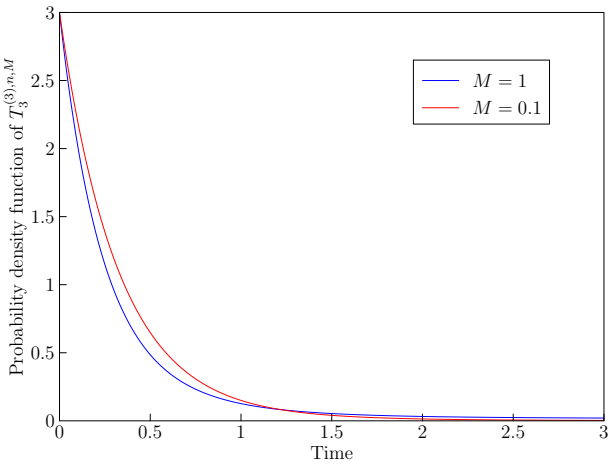

Figure 1: Plot of the probability density function of $T_3^{(3),n,M}$ for three genes sampled from the same deme (i.e. in sampling configuration $i = 1$) for a n-island model with $n = 10$ and $M = 1$ (blue), and $M = 0.1$ (red).

n-island model considered here, there always exists a panmictic model with population size changes that perfectly fits the T_2 distribution of this model.

A similar result can be obtained for the first coalescence time (denoted simply T_k) for a sample of size k, for any $k \geq 3$. However, the population size history mimicking the distribution of T_k is not exactly equal to the IICR, but to $\binom{k}{2}$ times the IICR. This comes from the fact that the first coalescence in a sample of size k arrives $\binom{k}{2}$ times faster than in a sample of size 2, because all pairs of lineages can coalesce. In particular, the population size function mimicking the distribution of the first coalescence time T_3, for a sample of three genes, is 3 times larger than the IICR corresponding to T_3, which can easily be seen by derivating and re-arranging equation (1) in order to obtain the corresponding IICR.

We study below this population size-change function for the symmetrical n-island model, and compare it asymptotically when $t \to \infty$ to that obtained by Mazet et al. [2016] for a sample of size 2. We will use the notation $\lambda(\cdot)$ for this population size-change function, rather than for the IICR. Another way of addressing this issue would be to define the $IICR_k = \binom{k}{2} \times IICR$. This notation issue did not arise in [Mazet et al., 2016], since in the particular case of $k = 2$ the two quantities are equal, namely $IICR_2 = IICR$.

Based on results from the previous section, the population size-change function mimicking the distribution of $T_3^{(3),n,M}$, for a sample of three genes in configuration $i = 1, 2, 3$, is given by

$$\lambda_i(t) = 3\,\frac{\mathbb{P}_i(T_3^{(3),n,M} > t)}{f_{T_3^{(3),n,M},i}(t)} = 3\,\frac{1 - P_t(i,4) - P_t(i,5)}{3P_t(i,1) + P_t(i,2)}. \tag{15}$$

In Figure 2 we plot this function for the different sampling schemes ($i = 1, 2, 3$), for a model with $n = 10$ islands and migration parameter $M = 1$ (left panel) and $M = 0.1$ (right panel). We can see that the three population size functions converge to a same asymptotic value when $t \to \infty$, which is confirmed by the following Proposition 5.

Proposition 5. *When $t \to \infty$, $\lambda_i(t)$, $i = 1, 2, 3$, have the following limit*

$$\lim_{t\to\infty} \lambda_i(t) = -\frac{3}{\mu_1}, \tag{16}$$

where μ_1 is the largest of the three distinct negative eigenvalues of the rate matrix Q.

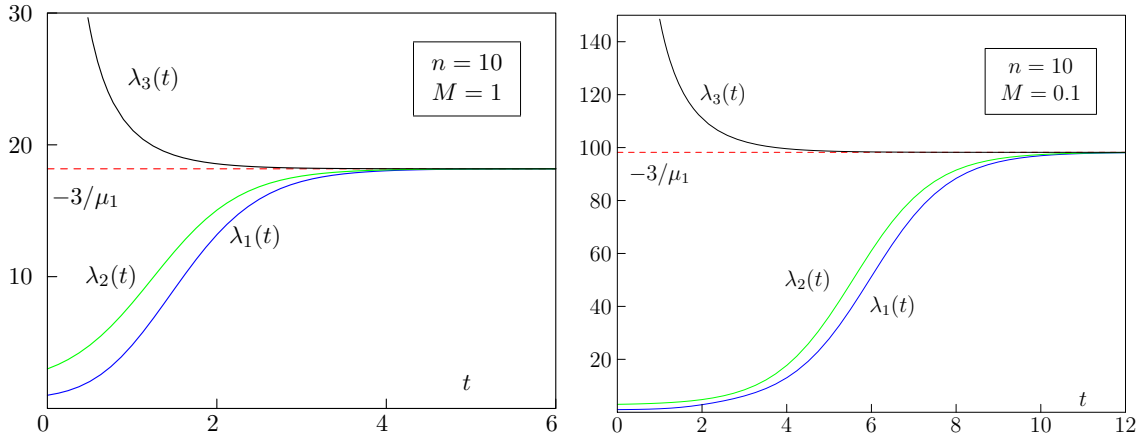

Figure 2: Plot of $\lambda_i(\cdot)$ $i = 1, 2, 3$ for $n = 10$ and $M = 1$ (left), and $M = 0.1$ (right). The dashed red line corresponds to the asymptotic value $-3/\mu_1$.

The proof is given in Appendix A.

This result is strikingly similar to that of Mazet et al. [2016] for a sample of size 2. Indeed, the population size-change function mimicking the distribution of T_2, for two genes sampled either from the same or from different demes, converges to $1/\beta$, where $-\beta$ is the largest of the two negative eigenvalues of the corresponding rate matrix (see Mazet et al. [2016]); an explicit formula for β is given in Equation (10)). As pointed out by Mazet et al. [2016], $1/\beta$ can be related to the notion of *effective population size* in a structured population. In particular, N/β is always strictly larger than the total population size nN, and for large M it is actually equal to the effective population size proposed by Nei and Takahata [1993]. For a sample of size 3, the inequality $-3/n < \mu_1$ in Lemma 1 implies that the asymptotic value of $N \times \lambda_i(\cdot)$ is also strictly larger than nN. However, as demonstrated below (Lemma 6 and Remark 7), this asymptotic value is also strictly larger than N/β. This highlights the ambiguity inherent to the notion of effective population size, as even this equilibrium value actually depends on the sample size used to compute it. See also Chikhi et al. [2018] for a discussion on several crucial differences between the IICR and the notion of N_e.

Lemma 6. *We have*

$$-3\alpha < \mu_3 < \mu_2 < \mu_1 < -3\beta < 0, \tag{17}$$

where $\mu_3 < \mu_2 < \mu_1 < 0$ are the ordered roots of $p(\mu)$ and $-\alpha < -\beta < 0$ are defined in Equation (10).

The proof is given in Appendix A.

Remark 7. *Together with Proposition 5, the result in Lemma 6 implies that, for $i = 1, 2, 3$, we have*

$$\lim_{t\to\infty} \lambda_i(t) > \frac{1}{\beta}.$$

Another similarity with the results of Mazet et al. [2016] is that $\lambda_i(\cdot)$ is increasing when all genes are sampled from the same island ($i = 1$), and decreasing when all genes are samples from different islands ($i = 3$) (see Figure 2). We do not provide a formal proof of this statement, only the following intuition: when all genes are sampled from the same island, the corresponding $IICR_3$ is initially lower than the asymptote, because the coalescence rate is greater in this state than in all other states of the ancestral lineage process. In contrast, when all genes are sampled

from different islands, coalescence is initially impossible because a migration has to occur first, so the $IICR_3$ is infinite. The last sampling configuration $i = 2$ (two genes sampled from the same island and the third one from a different island) does not exist for a sample of size 2, and is thus specific to the case $k = 3$. As expected intuitively, the curve of the corresponding $IICR_3$, $\lambda_2(\cdot)$, lies between the two other curves, and when the migration rate M is low the curve becomes increasingly closer to that of $\lambda_1(\cdot)$.

3 Using the joint distribution of coalescence times $(T_3^{(3)}, T_2^{(3)})$ to distinguish structure from panmixia

We will now compare the two models described in Subsections 2.1 and 2.2 : a panmictic model with population size-change function $\lambda(\cdot)$, and a symmetrical n-island model with parameters $n \geq 3$ and $M > 0$. As previously, we denote $T_3^{(3),\lambda}$ and $T_2^{(3),\lambda}$ the two coalescence times in the panmictic model, and $T_3^{(3),n,M}$ and $T_2^{(3),n,M}$ the two coalescence times in the n-island model.

We want to show that the joint distribution of $(T_3^{(3),\lambda}, T_2^{(3),\lambda})$ is always different from the joint distribution of $(T_3^{(3),n,M}, T_2^{(3),n,M})$, for all values of n, M, $\lambda(\cdot)$ and all initial sampling configurations ($i = 1, 2, 3$) for the n-island model.

In order to do so, we will fix n, M and a sampling configuration i, and consider the function $\lambda_i(\cdot)$ defined in Equation (15), for which the distributions of the first coalescence times $T_3^{(3),n,M}$ and $T_3^{(3),\lambda_i}$ are the same. We will show that even in this case, the distribution of $T_2^{(3),\lambda_i}$ is different from that of $T_2^{(3),n,M}$. This demonstration will be based on the study of the conditional distributions of $T_2^{(3)}$ given $T_3^{(3)}$ in the two models.

For several values of n and M, we will also quantify the difference between the two models by comparing the marginal distributions of the second coalescence times $T_2^{(3),n,M}$ and $T_2^{(3),\lambda_i}$, in the case when the distributions of the first coalescence times $T_3^{(3),n,M}$ and $T_3^{(3),\lambda_i}$ are the same.

3.1 Comparison of the conditional distributions of $T_2^{(3)}$ given $T_3^{(3)}$

We start by comparing, for a given $t > 0$ and a given sampling configuration $i = 1, 2, 3$, the conditional distribution functions $\mathbb{P}(T_2^{(3),\lambda_i} \leq \cdot | T_3^{(3),\lambda_i} = t)$ and $\mathbb{P}_i(T_2^{(3),n,M} \leq \cdot | T_3^{(3),n,M} = t)$. Using the results from Subsection 2.1, we deduce that

$$\begin{aligned}\mathbb{P}(T_2^{(3),\lambda_i} \leq u | T_3^{(3),\lambda_i} = t) &= 1 - \exp\{-(\Lambda_i(t+u) - \Lambda_i(t))\} \\ &= 1 - \left(\frac{1 - P_{t+u}(i,4) - P_{t+u}(i,5)}{1 - P_t(i,4) - P_t(i,5)}\right)^{1/3}. \end{aligned} \tag{18}$$

On the other hand, an expression for $\mathbb{P}_i(T_2^{(3),n,M} \leq \cdot | T_3^{(3),n,M} = t)$ was given in Equation (11).

In order to compare these two conditional distributions, let us introduce the functions $g_i(u,t)$, $i = 1, 2, 3$, defined for $u, t > 0$ by

$$g_i(u,t) := \mathbb{P}_i(T_2^{(3),n,M} \leq u | T_3^{(3),n,M} = t) - \mathbb{P}(T_2^{(3),\lambda_i} \leq u | T_3^{(3),\lambda_i} = t).$$

In Figures 3 and 4 we plot respectively the functions $g_1(u,t)$ and $g_2(u,t)$ as functions of u and t, in the case of a symmetrical n-island model with $n = 10$ and $M = 1$ (left panels), and $M = 0.1$ (right panels).

In these particular cases, we clearly see that there exists at least one pair (u,t) for which $g_i(u,t)$ is different from zero, *i.e.* for which the two conditional distributions are different. In order to demonstrate that this is the case for any choice of n, M and i, we further study in detail the behaviour of the functions $g_i(u,t)$ in the neighborhood of $(u,t) = (0,0)$ (Proposition 8) and their asymptotic behaviour when $u \to \infty$ (Proposition 9) or $t \to \infty$ (Proposition 10).

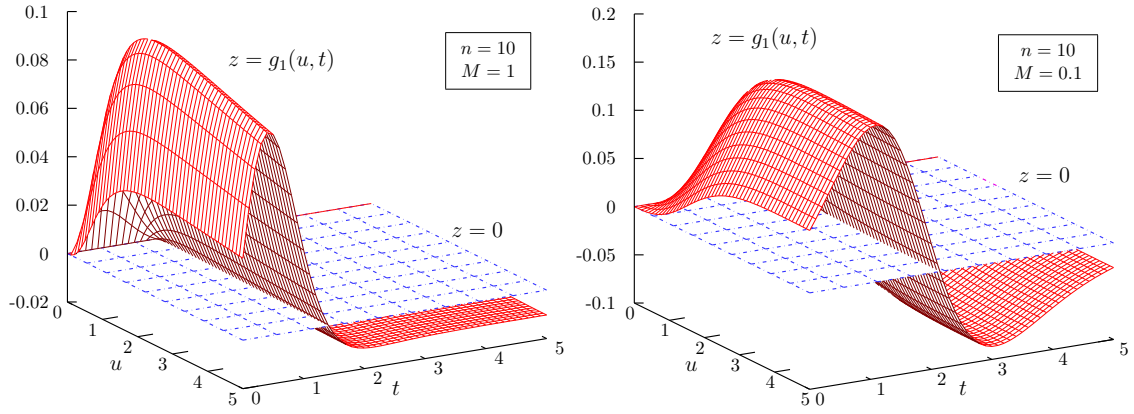

Figure 3: Difference of the conditional *cdf* in Equations (11) and (18) as a function of u and t, for $n = 10$ and $M = 1$ (left), and $M = 0.1$ (right) and three genes sampled from the same deme.

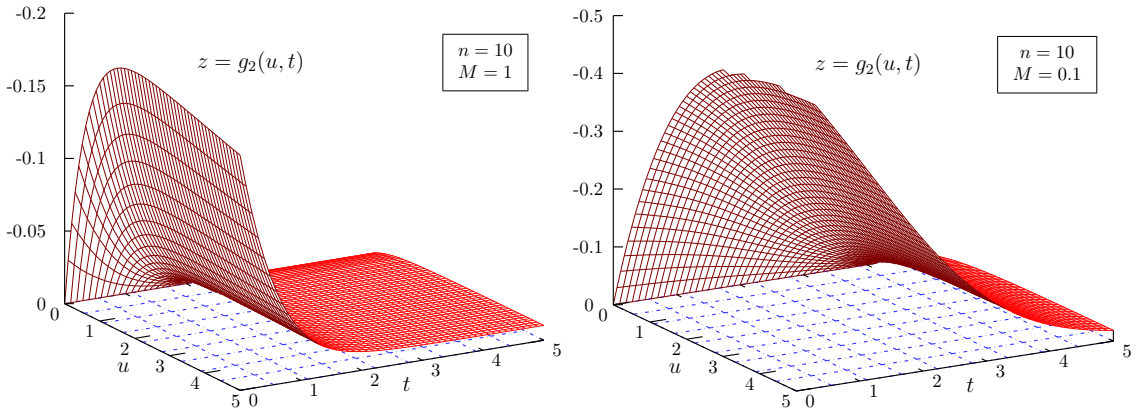

Figure 4: Difference of the conditional *cdf* in Equations (11) and (18) as a function of u and t, for $n = 10$ and $M = 1$ (left), and $M = 0.1$ (right) and three genes, two of which were sampled from the same deme and the third one from a different deme. Note that the z axis is inversed, for the sake of readability.

Proposition 8. *For (u,t) in the neighborhood of $(0,0)$ we have*

(i)

$$\begin{aligned}\mathbb{P}_1(T_2^{(3),n,M} \le u | T_3^{(3),n,M} = t) &= u - \frac{M+1}{2}u^2 - \frac{M}{2}ut + o(u^2+t^2),\\ \mathbb{P}_1(T_2^{(3),\lambda_1} \le u | T_3^{(3),\lambda_1} = t) &= u - \frac{M+1}{2}u^2 - Mut + o(u^2+t^2).\end{aligned}$$

As a consequence, we have $g_1(u,t) = \frac{M}{2}ut + o(u^2+t^2)$ as $(u,t) \to (0,0)$.

(ii)

$$\begin{aligned}\mathbb{P}_2(T_2^{(3),n,M} \le u | T_3^{(3),n,M} = t) &= \frac{M}{2(n-1)}u^2 + \frac{3M}{2(n-1)}ut + o(u^2+t^2),\\ \mathbb{P}_2(T_2^{(3),\lambda_2} \le u | T_3^{(3),\lambda_2} = t) &= \frac{1}{3}u - \frac{3M(n-3)+n-1}{18(n-1)}u^2\\ &\quad - \frac{(n-3)M}{3(n-1)}ut + o(u^2+t^2).\end{aligned}$$

Thus $g_2(u,t) = -\frac{u}{3} + \frac{3Mn+n-1}{18(n-1)}u^2 + \frac{M(2n+3)}{6(n-1)}ut + o(u^2+t^2)$ as $(u,t) \to (0,0)$.

(iii)

$$\begin{aligned}\mathbb{P}_3(T_2^{(3),n,M} \le u | T_3^{(3),n,M} = t) &= \frac{M}{2(n-1)}u^2 + \frac{3M}{4(n-1)}ut + o(u^2+t^2),\\ \mathbb{P}_3(T_2^{(3),\lambda_3} \le u | T_3^{(3),\lambda_3} = t) &= \frac{M}{2(n-1)}u^2 + \frac{M}{n-1}ut + o(u^2+t^2).\end{aligned}$$

Thus $g_3(u,t) = -\frac{M}{4(n-1)}ut + o(u^2+t^2)$ as $(u,t) \to (0,0)$.

Proposition 9. *For fixed $t > 0$, when $u \to +\infty$,*

$$g_i(u,t) = -K_{1,i}(n,M,t)e^{-\beta u} + o(e^{-\beta u}),\ i = 1,2,3, \tag{19}$$

where $K_{1,i}(n,M,t) > 0$ is given by

$$K_{1,i}(n,M,t) = \frac{3P_t(i,1)}{3P_t(i,1)+P_t(i,2)}\frac{1-a}{\beta} - \frac{P_t(i,2)}{3P_t(i,1)+P_t(i,2)}\frac{c}{\beta}, \tag{20}$$

with the constants β, a, c defined in Equations (9) and (10).

Proposition 10. *For fixed $u > 0$, we have*

$$\lim_{t\to+\infty} g_i(u,t) = -K_3(n,M,u),\ i = 1,2,3, \tag{21}$$

where $K_3(n,M,u) > 0$ is given by

$$K_3(n,M,u) = c_1 e^{-\beta u} + c_2 e^{-\alpha u} - e^{\frac{\mu_1}{3}u}, \tag{22}$$

with

$$\phi(\mu_1) = \frac{3M}{2(n-1)\mu_1 + 3Mn + 6(n-1)},\ c_1 = \frac{\phi(\mu_1)-\alpha}{\beta-\alpha},\ c_2 = \frac{\beta-\phi(\mu_1)}{\beta-\alpha}. \tag{23}$$

The proofs of these three propositions are given in Appendix A.

Propositions 8, 9 and 10 imply that for every $n \geq 3$ and $M > 0$, we can always find at least one pair (u, t) for which the two conditional distributions in Equations (11) and (18) are different. Indeed, such pairs can easily be exhibited for small values of both u and t, or for large values of u or t. As a consequence, the joint distribution of $(T_3^{(3),n,M}, T_2^{(3),n,M})$ for a sample of three genes is always different from the distribution of $(T_3^{(3),\lambda}, T_2^{(3),\lambda})$ in a panmictic population, for any possible size-change function $\lambda(\cdot)$ and any initial sampling configuration.

The results of this section also provide interesting insights into the comparison of coalescence times between panmictic and symmetrical n-island models. When the three genes are sampled from two or three different islands ($i = 2, 3$), $g_i(u, t)$ is negative at least in the neighboorhood of $(0, 0)$ and for sufficiently large values of u or t. This is in line with Figure 4, which suggests that $g_i(u, t)$ is always negative; if true, this implies that, for a panmictic and a symmetrical n-island model that have the same distribution of $T_3^{(3)}$, the *cdf* of $T_2^{(3)}$ is always smaller in the n-island model, which actually means that $T_2^{(3)}$ is stochastically larger in the n-island model than in the panmictic model.

The situation is more complex when the three genes are sampled from the same deme. For fixed $t > 0$ and sufficiently large values of u, $g_1(u, t)$ is also negative, meaning that, given that $T_3^{(3)}$ has the same value in the panmictic population and in the symmetrical n-island model, $T_2^{(3)}$ tends to be asymptotically larger in the n-island model than in the panmictic case. However, for small values of both u and t, $g_1(u, t)$ is positive, so for a sufficiently small common value $T_3^{(3)} = t$ and for small values of the second coalescence time $T_2^{(3)}$, this last coalescence time tends to be smaller in the n-island model than in the panmictic case. Actually, this latter situation is by far the most relevant in practice, because the asymptotics when $u \to \infty$ is reached very slowly (see Figure 3) and the probability density of $T_3^{(3)}$ is mostly concentrated on small values of t (see Figure 1).

Note also that, in the neighbourhood of $(0, 0)$, the leading term of $g_i(u, t)$ is very different depending on i: it is of order ut and depends linearly on M for $i = 1$ or $i = 3$, while it is of order u and independent of M for $i = 2$.

3.2 Comparison of the marginal distributions of $T_2^{(3)}$

The results of the previous section prove that the joint distributions of $(T_3^{(3),\lambda_i}, T_2^{(3),\lambda_i})$ and $(T_3^{(3),n,M}, T_2^{(3),n,M})$ always differ, even if the size-change function $\lambda_i(\cdot)$ is such that $T_3^{(3),\lambda_i}$ and $T_3^{(3),n,M}$ have the same distribution.

In order to quantify the distance between the coalescence time distributions in the panmictic case *versus* the symmetrical n-island model, it is easier in practice (even if less informative) to compare only the marginal distributions of $T_2^{(3),\lambda_i}$ and $T_2^{(3),n,M}$, in the case when the size-change function $\lambda_i(\cdot)$ is such that the distributions of $T_3^{(3),\lambda_i}$ and $T_3^{(3),n,M}$ are the same.

Using the results from the previous section, we have

$$\begin{aligned}\mathbb{P}(T_2^{(3),\lambda_i} \leq u) &= 1 - \int_0^\infty \mathbb{P}(T_2^{(3),\lambda_i} > u | T_3^{(3),\lambda_i} = t) f_{T_3^{(3),\lambda_i}}(t) dt \\ &= 1 - \int_0^\infty \left(\frac{1 - P_{t+u}(i,4) - P_{t+u}(i,5)}{1 - P_t(i,4) - P_t(i,5)} \right)^{1/3} [3P_t(i,1) + P_t(i,2)] dt,\end{aligned}$$

which we will compare with the expression for $\mathbb{P}_i(T_2^{(3),n,M} \leq u)$ provided by Equations (12) and (13).

Figure 5 shows these two marginal distribution functions for a symmetrical n-island model with $n = 10$ and $M = 1$ (left), and $M = 0.1$ (right), and three genes sampled in the same deme

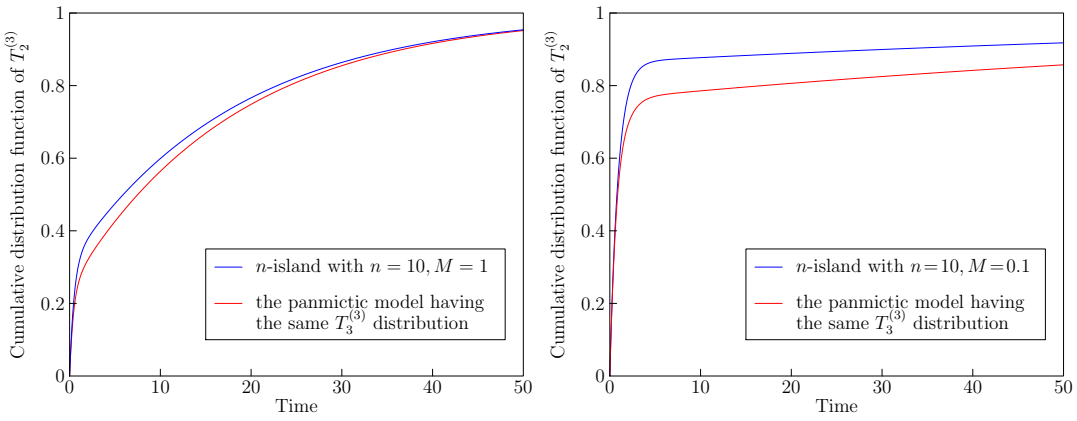

Figure 5: Comparison of the *cdf*s of $T_2^{(3),\lambda_1}$ and $T_2^{(3),n,M}$ for three genes sampled in the same deme, for a n-island model with $n = 10$ and $M = 1$ (left), and $M = 0.1$ (right).

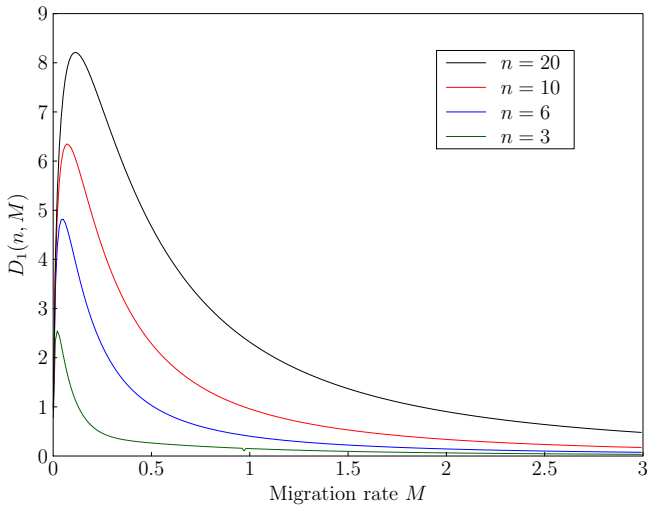

Figure 6: 1-Wasserstein distance $D_1(n, M)$ between the distributions of $T_2^{(3),n,M}$ and $T_2^{(3),\lambda_1}$ as a function of M, for different values of n.

(i.e. in sample configuration $i = 1$). As discussed in the previous section, for this particular sampling configuration, the *cdf* of $T_2^{(3)}$ appears to be larger (hence $T_2^{(3)}$ stochastically smaller) in the n-island model than in the panmictic case.

We compare the marginal probability distributions of $T_2^{(3),n,M}$ and $T_2^{(3),\lambda_i}$ using the 1-Wasserstein distance (see *e.g.* Vallander [1973]), which we denote

$$\begin{aligned} D_i(n, M) :&= \int_0^\infty \left| \mathbb{P}_i(T_2^{(3),n,M} \leq u) - \mathbb{P}(T_2^{(3),\lambda_i} \leq u) \right| du \\ &= \int_0^\infty \left| \int_0^\infty g_i(u,t)[3P_t(i,1) + P_t(i,2)]dt \right| du. \end{aligned}$$

In Figure 6 we show several plots of $D_1(n, M)$ as a function of M, for different values of n, in the case of three genes sampled in the same deme ($i = 1$). We see that, as expected, the difference between the two distributions increases with n, the number of demes, and decreases when M is large or very small. When M increases the n-island model tends towards a large panmictic population of size nN. When M is very small, each deme is nearly isolated from the other demes, and thus increasingly behaves as a simple panmictic population of size N.

4 Discussion and perspectives

In this study we obtain new theoretical results about the joint distribution of the coalescence times $(T_3^{(3)}, T_2^{(3)})$ for a sample of three genes in a symmetrical n-island model with constant size, for all possible sampling configurations. When comparing this distribution with the analogous one in a panmictic population with population size changes, we show that for any size-change function $\lambda(\cdot)$, the two distributions are different. Indeed, it is always possible to construct a size-change function $\lambda(\cdot)$ which perfectly mimics the distribution of $T_3^{(3)}$ in a symmetrical n-island model; but we show that even in this case, the conditional distributions of $T_2^{(3)}$ given $T_3^{(3)}$ in this panmictic model and in the n-island model are different. Consequently, our results imply that the joint distribution of coalescence times for a sample of three genes contains enough information to distinguish between a panmictic population and a symmetrical n-island model of constant size. As illustrated by Figures 5 and 6, the difference between the two models can be substantial, even in the case where these models lead to the same distribution of $T_3^{(3)}$.

Although this result is extremely important from a theoretical perspective, we note that, currently, it cannot be directly implemented into a model choice procedure based on real genomic data. Indeed, to our knowledge, no statistical method has yet been proposed to estimate neither the joint distribution of $(T_3^{(3)}, T_2^{(3)})$ or the marginal distribution of $T_2^{(3)}$ from genomic data. However, some important progresses have been made recently into this direction. For instance, the distribution of the first coalescence time T_k in a sample of size k can be estimated quite accurately by the MSMC approach [Schiffels and Durbin, 2014], and other approaches might be implemented in the near future to estimate also more ancient coalescent times (see for instance the recent work of Weissman and Hallatschek [2017]). Besides, the fact that considering jointly several coalescence times allows discriminating structured and panmictic models, might be exploited by statistical procedures based on the Site Frequency Spectrum (SFS). Indeed, this quantity is directly observable from genomic data, as it records the proportion of genomic positions with $1, 2, \dots n-1$ copies of the mutant allele in a sample of n genes, and the limit of this quantity for large numbers of loci can be expressed as a linear function of the successive expected coalescence times $\mathbb{E}(T_n), \mathbb{E}(T_{n-1}), \dots, \mathbb{E}(T_2)$ [Griffiths and Tavaré, 1998].

Moreover, it should be possible in theory to apply the ideas presented in this article to compare the joint distribution of $(T_3^{(3)}, T_2^{(3)})$ between a panmictic population and a structured population for more general models of population structure. In this article we focused on the symmetrical n-island model, but our results can be straightforwardly generalized to any model of population structure, provided that we can compute the corresponding transition kernel P_t. The theoretical results of Herbots (see Herbots [1994], Wilkinson-Herbots [1998]) allow one to write the rate matrix Q for any given model of population structure under the *structured coalescent*. One important issue is that complex models are characterized by very large rate matrices Q, on the order of $n^2 \times n^2$ for $k = 2$. For a model with n demes, it may be difficult to study the transition kernel P_t, when the rate matrix Q has on the order of 10,000 elements. This is currently explored by Rodriguez et al. [201X] who show that the $IICR_k$ can be obtained for highly complex models. Altogether, the formal proofs presented here and the work of Rodriguez et al. [201X] suggest that demographic inference under complex models of population structure may become easier in the next few years.

Appendix A. Proofs for Sections 2 and 3

Proof of Lemma 1

We first observe that

$$-2 - \frac{nM}{n-1} < -\frac{3}{n} < 0.$$

Then the calculus gives

$$\begin{aligned} p(0) &= \frac{9M(Mn+2n-2)}{2(n-1)^2} > 0, \\ p\left(-\frac{3}{n}\right) &= -\frac{9(Mn^2+2(n-1)(n-3))}{2n^3} < 0, \\ p\left(-2-\frac{nM}{n-1}\right) &= \frac{(Mn+2n-2)(M(n-3)+2n-2)}{2(n-1)^2} > 0 \\ \lim_{\mu\to-\infty} p(\mu) &= -\infty. \end{aligned}$$

The intermediate value theorem applied to $p(\mu)$ thus provides a proof of the result. □

Proof of Proposition 3

In order to simplify computations, we make the following change of parameters:

$$M = 2N,\ n = \frac{a+N}{a}.$$

The new parameters a et N verify $a > 0$ and $N > 0$ and the matrix Q becomes

$$Q = \begin{bmatrix} -3-3\,N & 3\,N & 0 & 3 & 0 \\ a & -1+a-2\,N & -2\,a+2\,N & 0 & 1 \\ 0 & 6\,a & -6\,a & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The matrix Q has the double eigenvalue 0 and it is easy to check that the corresponding eigenspace is generated by the following non colinear vectors:

$$V_1 = [N, N-1, N-1, 2N, N-a-1]^T,\ \ V_2 = [1,1,1,1,1]^T.$$

We then consider the change of basis matrix

$$P_1 = \begin{bmatrix} 1 & 0 & 0 & N & 1 \\ 0 & 1 & 0 & N-1 & 1 \\ 0 & 0 & 1 & N-1 & 1 \\ 0 & 0 & 0 & 2\,N & 1 \\ 0 & 0 & 0 & N-a-1 & 1 \end{bmatrix},$$

whose inverse is

$$P_1^{-1} = \begin{bmatrix} 1 & 0 & 0 & -\frac{a+1}{N+a+1} & -\frac{N}{N+a+1} \\ 0 & 1 & 0 & -\frac{a}{N+a+1} & -\frac{N+1}{N+a+1} \\ 0 & 0 & 1 & -\frac{a}{N+a+1} & -\frac{N+1}{N+a+1} \\ 0 & 0 & 0 & \frac{1}{N+a+1} & -\frac{1}{N+a+1} \\ 0 & 0 & 0 & -\frac{N-a-1}{N+a+1} & \frac{2N}{N+a+1} \end{bmatrix}.$$

We have thus obtained a partial diagonalization (by blocks) of the matrix Q:

$$Q_2 = P_1^{-1} Q P_1 = \begin{bmatrix} -3-3\,N & 3\,N & 0 & 0 & 0 \\ a & -1+a-2\,N & -2\,a+2\,N & 0 & 0 \\ 0 & 6\,a & -6\,a & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{24}$$

and we put

$$R := \begin{bmatrix} -3-3\,N & 3\,N & 0 \\ a & -1+a-2\,N & -2\,a+2\,N \\ 0 & 6\,a & -6\,a \end{bmatrix}.$$

The characteristic polynomial of R is the polynomial $p(\mu)$, which has the following expressions using the new parameters:

$$p(\mu) = \mu^3 + (5N+5a+4)\mu^2 + 3(2N^2+4aN+3N+2a^2+7a+1)\mu + 18a(N+a+1),$$

which has three strictly negative real roots. These roots are all distinct by Lemma 1.

If μ is an eigenvalue of R, we then determine a corresponding eigenvector $W(\mu)$ by solving the equation $RW(\mu) = \mu W(\mu)$.

The computations show that we can choose

$$W(\mu) = \begin{bmatrix} \mu^2 + (2N+5a+1)\mu + 6a(a+1) \\ a(\mu+6a) \\ 6a^2 \end{bmatrix}.$$

We then consider the 3×3 passage matrix P_2, whose column vectors are the $W(\mu_i), i = 1, 2, 3$, where the μ_i are the three eigenvalues of R:

$$P_2 = [W(\mu_1),\ W(\mu_2),\ W(\mu_3)]\,.$$

Some easy computations on the rows show that the determinant of P_2 is a Van der Monde determinant, and hence: $\det(P_2) = (\mu_1-\mu_2)(\mu_1-\mu_3)(\mu_2-\mu_3) \neq 0$.

The computations of the inverse P_2^{-1} gives

$$P_2^{-1} = [Z(\mu_1), Z(\mu_2), Z(\mu_3)]^T\,,$$

where

$$Z(\mu) = \frac{1}{p'(\mu)} \begin{bmatrix} 1 \\ \dfrac{\mu+3N+3}{a} \\ -\dfrac{\mu^2+(3N+a+4)\mu+3(N+a+1)}{a\mu} \end{bmatrix},$$

with

$$p'(\mu) = 3\mu^2 + 2(5N+5a+4)\mu + 3(2N^2+4aN+3N+2a^2+7a+1).$$

We further obtain

$$\begin{aligned} e^{tR} &= P_2 \begin{bmatrix} e^{\mu_1 t} & 0 & 0 \\ 0 & e^{\mu_2 t} & 0 \\ 0 & 0 & e^{\mu_3 t} \end{bmatrix} P_2^{-1}, \\ &= \sum_{j=1}^{3} e^{\mu_j} \; W(\mu_j) Z(\mu_j)^T. \end{aligned}$$

We then introduce the matrices $A(\mu)$ and B defined by

$$A(\mu) := P_1 \; \bar{W}(\mu) \bar{Z}(\mu)^T \; P_1^{-1},$$

where $\bar{W}_j(\mu)$ (resp. $\bar{Z}(\mu)$) is a vector of length 5 obtained by adding two null coordinates to $W(\mu)$ (resp. $Z(\mu)$), and

$$B = P_1 \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} P_1^{-1}.$$

Emphasizing the rank one property of matrix $A(\mu)$, let us define the vectors

$$E_1 = \begin{bmatrix} \dfrac{3M}{v} \\ 1 \\ \dfrac{3M}{u} \\ 0 \\ 0 \end{bmatrix}, \quad E_2 = \begin{bmatrix} Mu \\ (n-1)uv \\ (n-1)(n-2)Mv \\ \dfrac{3Mu}{\mu} \\ \dfrac{(n-1)uv}{\mu} \end{bmatrix},$$

where μ is an eigenvalue of R, and u, v are functions of μ defined by

$$u = (n-1)\mu + 3M, \quad v = 2\mu + 3(M+2).$$

Note that, since

$$p\left(-\frac{3M}{n-1}\right) = -\frac{9}{2} \cdot \frac{(n-2)[2(n-1)+(n-3)M]M^2}{(n-1)^3} < 0,$$

$$p\left(-\frac{3(M+2)}{2}\right) = \frac{9}{8} \cdot \frac{[2(n-1)+(n-3)M]M^2}{(n-1)^2} > 0,$$

$-\dfrac{3M}{n-1}$ and $-\dfrac{3(M+2)}{2}$ cannot be eigenvalues and thus $u \neq 0$ et $v \neq 0$.

With $\delta(M, n, \mu) = 2(n-1)^2 p'(\mu)$, we have

$$A(\mu) = \frac{1}{\delta(M, n, \mu)} \; E_1 E_2^T, \tag{25}$$

which gives the expression of $A(\mu)$ in Proposition 3.

The stated result easily follows. □

Proof of Proposition 5

Using Equation (15) and Proposition 3, we deduce that

$$\lim_{t\to\infty} \lambda_i(t) = -3\ \frac{A(\mu_1)(i,4) + A(\mu_1)(i,5)}{3A(\mu_1)(i,1) + A(\mu_1)(i,2)},\ i = 1,2,3.$$

Note that the matrix $E_1E_2^T$ introduced in the proof of Proposition 3 can also be factorized in a different manner. If we define the 3×3 matrix C, the 5×3 matrix D_1 and the 3×5 matrix D_2 by

$$C := \begin{bmatrix} \frac{u}{v} & u & 1 \\ u & uv & v \\ 1 & v & \frac{v}{u} \end{bmatrix},\quad D_1 := \begin{bmatrix} 3M & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3M \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$D_2 := \begin{bmatrix} M & 0 & 0 & \frac{3M}{\mu} & 0 \\ 0 & n-1 & 0 & 0 & \frac{n-1}{\mu} \\ 0 & 0 & (n-1)(n-2) & 0 & 0 \end{bmatrix},$$

we may check that, for any value of μ, we have $E_1E_2^T = D_1CD_2$.

The vectors V_1 and V_2, defined by

$$V_1 := \begin{bmatrix} 3 \\ 0 \\ 0 \\ -\mu \\ 0 \end{bmatrix},\quad V_2 := \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ -\mu \end{bmatrix},$$

verify $D_2V_j = 0$ for $j = 1,2$, and hence, using Equation (25), $A(\mu)V_j = 0$, for $j = 1,2$.

Therefore, for $i = 1,2,3$,

$$3A(\mu)(i,1) - \mu A(\mu)(i,4) = 0,\ A(\mu)(i,2) - \mu A(\mu)(i,5) = 0,$$

and the result follows. □

Proof of Lemma 6

Note that -3α and -3β, with α, β defined in Equation (10), are the roots of the polynomial $q_1(X) = q(X/3)$, where

$$q(X) = X^2 + \left(1 + \frac{nM}{n-1}\right) X + \frac{M}{n-1},$$

and on the open subset $D = \{(n,M) :\ n > 2,\ M > 0\}$ of $\mathbb{R}^2$ the polynomials $p(X)$ and $q_1(X)$ have a common root if and only their resultant $R(n,M) = \mathrm{Res}(p(X), q_1(X), X)$ with respect to X is null (see e.g. [Lang, 2002, chapter IV-8]).

Because $R(n,M) = -\frac{M^3}{9(n-1)^2} < 0$ on D, and because the roots -3α and μ_3 are continuous functions of (n,M) on D, the inequality $-3\alpha < \mu_3$ for one value of $(n,M) \in D$ implies the inequality everywhere on D. The same is true for the inequality $\mu_1 < -3\beta$.

This achieves the proof of the lemma. □

Proof of Proposition 8

We will only give the proof of (i), which corresponds to the case of three genes sampled from the same deme. The proofs ot the analougous results (ii) and (iii), corresponding to the two other sample schemes, are similar.

When $t \to 0$, using the well-known relation $P_t = I_5 + tQ + o(t)$, where I_5 denotes the identity matrix, we have the following Taylor expansions :

$$P_t(1,1) = 1 + \left(-\frac{3M}{2} - 3\right) t + o(t),$$
$$P_t(1,2) = \frac{3M}{2} t + o(t).$$

In particular, we have

$$\frac{1}{3P_t(1,1) + P_t(1,2)} = \frac{1}{1 - (M+3)t + o(t)} = 1 + (M+3)t + o(t).$$

Using Equations (7), (8), (9), (10) and (11), and a Taylor expansion of order 2 of the exponential in the neighborhood of 0, we easily obtain

$$F_{2,s}(u) = u - \frac{M+1}{2} u^2 + o(u^2),$$
$$F_{2,d}(u) = \frac{M}{2(n-1)} u^2 + o(u^2).$$

When substituting the above expressions into Equation (11), straightforward calculations give

$$\mathbb{P}_1(T_2^{(3),n,M} \leq u | T_3^{(3),n,M} = t) = u - \frac{M}{2} ut - \frac{M+1}{2} u^2 + o(u^2 + t^2).$$

Further, using Equation (18), let us denote

$$h(u,t) := \mathbb{P}(T_2^{(3),\lambda_1} \leq u | T_3^{(3),\lambda_1} = t) = 1 - \left(\frac{1 - P_{t+u}(1,4) - P_{t+u}(1,5)}{1 - P_t(1,4) - P_t(1,5)} \right)^{1/3}.$$

We will write a Taylor expansion of order 2 of $h(u,t)$ in the neighborhood of $(0,0)$. We have $h(0,0) = 0$ and direct computations give

$$\begin{aligned}
\frac{\partial h}{\partial u}(u,t) &= \frac{1}{3} \times \left(1 - P_t(1,4) - P_t(1,5)\right)^{-1/3} \times \left(1 - P_{t+u}(1,4) - P_{t+u}(1,5)\right)^{-2/3} \\
&\quad \times \left(P_{t+u}Q(1,4) + P_{t+u}Q(1,5)\right), \\
\frac{\partial h}{\partial t}(u,t) &= \frac{\left(1 - P_{t+u}(1,4) - P_{t+u}(1,5)\right)^{-2/3}}{3\left(1 - P_t(1,4) - P_t(1,5)\right)^{4/3}} \\
&\quad \times \{\left(P_{t+u}Q(1,4) + P_{t+u}Q(1,5)\right)\left(1 - P_t(1,4) - P_t(1,5)\right) \\
&\quad - \left(P_tQ(1,4) + P_tQ(1,5)\right)\left(1 - P_{t+u}(1,4) - P_{t+u}(1,5)\right)\}.
\end{aligned}$$

Using the fact that $P_0 = I_5$, and $Q(1,4) = 3, Q(1,5) = 0$, we obtain

$$\frac{\partial h}{\partial u}(0,0) = 1, \frac{\partial h}{\partial t}(0,0) = 0.$$

We further compute the second partial derivatives of h in $(0,0)$. With Q^2 being the square matrix of the rate matrix Q, and using the fact that $Q^2(1,4) = -9\left(\frac{M}{2} + 1\right)$ and $Q^2(1,5) = \frac{3M}{2}$,

we easily derive

$$\frac{\partial^2 h}{\partial u \partial t}(0,0) = \frac{1}{3}\left\{(Q(1,4)+Q(1,5))^2 + Q^2(1,4) + Q^2(1,5)\right\} = -M,$$
$$\frac{\partial^2 h}{\partial u^2}(0,0) = \frac{1}{3}\left\{\frac{2}{3}(Q(1,4)+Q(1,5))^2 + Q^2(1,4) + Q^2(1,5)\right\} = -(M+1),$$
$$\frac{\partial^2 h}{\partial t^2}(0,0) = 0,$$

The Taylor expansion of order 2 of $h(u,t)$ near $(0,0)$ finally gives

$$\mathbb{P}(T_2^{(3),\lambda_1} \leq u | T_3^{(3),\lambda_1} = t) = u - Mut - \frac{M+1}{2}u^2 + o(u^2+t^2),$$

which finishes the proof. □

Proof of Proposition 9

Because $0 < \beta < \alpha$, using Equations (7) and (8), we get

$$F_{2,s}(u) = 1 - \frac{a}{\alpha}e^{-\alpha u} - \frac{1-a}{\beta}e^{-\beta u} = 1 - \frac{1-a}{\beta}e^{-\beta u} + o(e^{-\beta u})$$

and

$$F_{2,d}(u) = 1 - \frac{c}{\alpha}e^{-\alpha u} + \frac{c}{\beta}e^{-\beta u} = 1 + \frac{c}{\beta}e^{-\beta u} + o(e^{-\beta u}).$$

Therefore, using Equation (11), we obtain

$$\mathbb{P}_i(T_2^{(3),n,M} \leq u | T_3^{(3),n,M} = t) = 1 - K_{1,i}(n,M,t)e^{-\beta u} + o(e^{-\beta u}),$$

where $K_{1,i}(n,M,t)$, given by (20), is strictly positive because $0 < a < 1$ and $c < 0$.

Using Proposition 3 we get

$$\sum_{j=1}^{3} P_{t+u}(i,j) = \left(A(\mu_1)(i,1) + A(\mu_1)(i,2) + A(\mu_1)(i,3)\right) e^{\mu_1(t+u)} + o(e^{\mu_1(t+u)}).$$

Therefore, using Equation (18) and relation $1 - P_t(i,4) - P_t(i,5) = \sum_{j=1}^{3} P_t(i,j)$, we obtain

$$\mathbb{P}(T_2^{(3),\lambda_i} \leq u | T_3^{(3),\lambda_i} = t) = 1 - K_{2,i}(n,M,t)\, e^{\frac{\mu_1}{3}u} + o(e^{\frac{\mu_1}{3}u}),$$

where

$$K_{2,i}(n,M,t) := \left(\frac{A(\mu_1)(i,1) + A(\mu_1)(i,2) + A(\mu_1)(i,3)}{P_t(i,1) + P_t(i,2) + P_t(i,3)}\right)^{\frac{1}{3}} e^{\frac{\mu_1}{3}t}.$$

Because $\frac{\mu_1}{3} < -\beta$ from Lemma 6, we have $e^{\frac{\mu_1}{3}u} = o(e^{-\beta u})$, which achieves the proof of (19). □

Proof of Proposition 10

We will first prove a useful lemma.

Lemma 11. *Let us define $\phi(\mu)$ by*

$$\phi(\mu) := \frac{3A(\mu)(1,1)}{3A(\mu)(1,1) + A(\mu)(1,2)},$$

where the matrix $A(\mu)$ is given in Proposition 3.

Then,

(i) $0 < \phi(\mu_i) < -\frac{\mu_i}{3} < \alpha,\ i = 1, 2, 3.$

(ii) Defining $h(n)$ *by*

$$h(n) = \frac{(n-1)(25 - 9n + 5\sqrt{9n^2 - 18n + 25})}{12n^2},$$

we have the following results:

(a) If $M > h(n)$*, then* $\phi(\mu_1) < \beta < \phi(\mu_2) < \phi(\mu_3) < \alpha.$

(b) If $M = h(n)$*, then* $\phi(\mu_1) < \beta = \phi(\mu_2) < \phi(\mu_3) < \alpha.$

(c) If $M < h(n)$*, then* $\phi(\mu_1) < \phi(\mu_2) < \beta < \phi(\mu_3) < \alpha.$

Proof. From Proposition 3 and using $p(\mu_i) = 0$, we get

$$\begin{aligned}\phi(\mu_i) &= \frac{1}{3}\frac{(n-1)\mu_i^2 + (nM + n - 1)\mu_i + 6M}{Mn + 2(n-1)} \\ &= \frac{3(n-1)}{nM + 2(n-1)}\, q\left(\frac{\mu_i}{3}\right) - \frac{M(\mu_i n + 3)}{nM + 2(n-1)} - \frac{\mu_i}{3}.\end{aligned}$$

From Lemma 1, it follows that $\mu n + 3 > 0$. Because $-\alpha < \dfrac{\mu_i}{3} < -\beta$ from Lemma 6, we get $q\left(\frac{\mu_i}{3}\right) < 0$, that proves $\phi(\mu_i) < -\frac{\mu_i}{3}$.

Now, using again that $p(\mu_i) = 0$, we obtain the following expression for $\phi(\mu_i)$

$$\phi(\mu_i) = \frac{3M}{2(n-1)\mu_i + 3nM + 6(n-1)}. \tag{26}$$

Introducing the new variable $\nu = \dfrac{3M}{2(n-1)\mu + 3nM + 6(n-1)}$, and thus $\mu = -\dfrac{3}{2}\dfrac{(nM + 2n - 2))\nu - M}{(n-1)\nu}$, we get that $\phi(\mu_i)$, for $i = 1, 2, 3$ are the three real roots of the polynomial

$$\begin{aligned}r(\nu) &= 4(n-1)(nM + 2n - 2)\nu^3 - [n^2M^2 + 2(n-1)(3n+4)M + 8(n-1)^2]\nu^2 \\ &\quad +2M(2nM + 5n - 5)\nu - 3M^2.\end{aligned}$$

The coefficient signs of $r(X)$ show that $r(X)$ has no negative roots, implying that $\phi(\mu_i) > 0,\ i = 1, 2, 3$, which achieves the proof of (i).

The set of (n, M) such that $r(X)$ and $q(-X)$ have a common root is obtained by computing their resultant with respect to X, denoted by $R(n, M)$:

$$R(n, M) = \mathrm{Res}(r(X), q(-X), X) = M^3\left[6n^2M^2 + (n-1)(9n-25)M - 6(n-1)^2\right].$$

In the domain $D = \{(n, M),\ n > 2,\ M > 0\}$ the curve $6n^2M^2 + (n-1)(9n-25)M - 6(n-1)^2 = 0$ is identical to the graph of the function $M = h(n),\ n > 2$.

The set $D \setminus \{(n, h(n)) :\ n > 2\}$ has two connex open components in which the relative position of β, α and $\phi(\mu_i),\ i = 1, 2, 3$ are the same.

In the component $D_1 := \{(n, M) :\ n > 2,\ M > h(n)\}$, one may choose $n = 3,\ M = \frac{4}{3}$ for which

$$\begin{aligned}
\beta &= \frac{3}{2} - \frac{\sqrt{57}}{6} \approx 0.241694, \\
\alpha &= \frac{3}{2} + \frac{\sqrt{57}}{6} \approx 2.758305, \\
\phi(\mu_1) &= \frac{3-\sqrt{5}}{4} \approx 0.190983, \\
\phi(\mu_2) &= \frac{1}{3}, \\
\phi(\mu_3) &= \frac{3+\sqrt{5}}{4} \approx 1.3090167,
\end{aligned}$$

that proves $(ii-a)$.

In the component $D_2 := \{(n, M) :\ n > 2,\ M < h(n)\}$, one may choose $n = 3,\ M = \frac{1}{2}$ for which

$$\begin{aligned}
\beta &= \frac{7-\sqrt{33}}{8} \approx 0.156929, \\
\alpha &= \frac{7+\sqrt{33}}{8} \approx 1.593070, \\
\phi(\mu_1) &\approx 0.104089, \\
\phi(\mu_2) &\approx 0.146359, \\
\phi(\mu_3) &\approx 1.118868,
\end{aligned}$$

that proves $(ii-c)$.

The curve arc $\{(n, h(n)) : n > 2\}$ may be parametrized by

$$n = \frac{3u^2+8u-3}{3(u^2-1)},\ M = \frac{8u}{(u+3)^2},\ u \in]1, 3[,$$

and we obtain

$$\begin{aligned}
\beta &= \frac{u-1}{u+3}, \\
\alpha &= \frac{3(u+1)}{u+3}, \\
\phi(\mu_1) &= \frac{39u+51-3\sqrt{-71u^2+442u+529}}{40(u+3)}, \\
\phi(\mu_2) &= \frac{u-1}{u+3}, \\
\phi(\mu_3) &= \frac{39u+51+3\sqrt{-71u^2+442u+529}}{40(u+3)},
\end{aligned}$$

and $(ii-b)$ is proved. □

Lemma 11 is now used to prove Proposition 10. First note that we have

$$\phi(\mu) = \frac{3A(\mu)(i,1)}{3A(\mu)(i,1) + A(\mu)(i,2)},$$

for every $i = 1, 2, 3$.

When $t \to +\infty$, $P_t(i,j) = A(\mu_1)(i,j)e^{\mu_1 t} + o(e^{\mu_1 t}),\ j = 1, 2$ and thus

$$\begin{aligned}
&\mathbb{P}_i(T_2^{(3),n,M} \leq u | T_3^{(3),n,M} = t) \\
&= \frac{3A(\mu_1)(i,1)}{3A(\mu_1)(i,1) + A(\mu_1)(i,2)} F_{2,s}(u) + \frac{A(\mu_1)(i,2)}{3A(\mu_1)(i,1) + A(\mu_1)(i,2)} F_{2,d}(u) + o(1) \\
&= \phi(\mu_1) F_{2,s}(u) + (1 - \phi(\mu_1)) F_{2,d}(u) + o(1).
\end{aligned}$$

Using the definitions of $F_{2,s}(u)$ and $F_{2,d}(u)$ in Equations (7) and (8), we get

$$\mathbb{P}_i(T_2^{(3),n,M} \leq u | T_3^{(3),n,M} = t) = 1 - c_1 e^{-\beta u} - c_2 e^{-\alpha u} + o(1),$$

where c_1 and c_2 are given by (23).

On another hand, we have $P_t(i,j) = A(\mu_1)(i,j)e^{\mu_1 t} + o(e^{\mu_1 t}),\ j = 1, 2, 3$, we obtain

$$\begin{aligned}
\sum_{j=1}^{3} P_{t+u}(i,j) &= \left((A(\mu_1)(i,1) + A(\mu_1)(i,2) + A(\mu_1)(i,3)\right) e^{\mu_1 (t+u)} + o(e^{\mu_1 t}), \\
\sum_{j=1}^{3} P_t(i,j) &= \left((A(\mu_1)(i,1) + A(\mu_1)(i,2) + A(\mu_1)(i,3)\right) e^{\mu_1 t} + o(e^{\mu_1 t}).
\end{aligned}$$

Using the fact that $1 - P_t(i,4) - P_t(i,5) = \sum_{j=1}^{3} P_t(i,j)$, this implies

$$\mathbb{P}(T_2^{(3),\lambda_i} \leq u | T_3^{(3),\lambda_i} = t) = 1 - e^{\frac{\mu_1}{3} u} + o(1),$$

and thus (21) holds.

It remains to show that $K_3(n, M, u) > 0$.

Using $c_1 + c_2 = 1$, we may write

$$K_3(n, M, u) = \frac{\beta - \phi(\mu_1)}{\beta - \alpha} \left(e^{-\alpha u} - e^{\frac{\mu_1}{3} u} \right) + \frac{\phi(\mu_1) - \alpha}{\beta - \alpha} \left(e^{\frac{\mu_1}{3} u} - e^{-\beta u} \right).$$

From Lemma 6, $-\alpha < \frac{\mu_1}{3} < -\beta$ and thus $e^{-\alpha u} < e^{\frac{\mu_1}{3} u} < e^{-\beta u}$. On the other hand, from Lemma 11 we have $\phi(\mu_1) < \beta < \alpha$. Therefore $K_3(n, M, u) > 0$. □

Appendix B. The case of $n = 2$ islands

If we now consider the ancestral lineage process for a sample of three genes in the case of a symmetrical n-island model with $n = 2$ two islands, we only have the following four possible configurations:

1. the three lineages are in the same island,
2. two lineages are in the same island and the third one is in the other island,
3. there are only two ancestral lineages left and they are in the same island,
4. there are only two ancestral lineages left and they are in different islands.

The corresponding transition rate matrix is:

$$Q = \begin{pmatrix} -\frac{3M}{2} - 3 & \frac{3M}{2} & 3 & 0 \\ \frac{M}{2} & -\frac{M}{2} - 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

The characteristic polynomial of Q is $\chi_Q(\mu) = -\mu^2 p(\mu)$, with

$$p(\mu) = \mu^2 + 2(M+2)\mu + 3(M+1).$$

The matrix Q has the double eigenvalue 0 and the corresponding eigenspace of dimension 2 can be generated by the vectors $\left[\frac{M}{2}, \frac{M}{2} - 1, M, -1\right]^T$ and $[1,1,1,1]^T$. The two other eigenvalues are $\mu_1 = -M - 2 + \sqrt{M^2 + M + 1}$ and $\mu_2 = -M - 2 - \sqrt{M^2 + M + 1}$.

An eigenvector for μ_1 (resp. μ_2) is $[3M, 2\mu_1 + 3M + 6, 0, 0]^T$ (resp. $[3M, 2\mu_2 + 3M + 6, 0, 0]^T$), and we may consider the following change of basis matrix P given by

$$P = \begin{pmatrix} 3M & 3M & \frac{M}{2} & 1 \\ 2\mu_1 + 3M + 6 & 2\mu_2 + 3M + 6 & \frac{M}{2} - 1 & 1 \\ 0 & 0 & M & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix}.$$

From

$$P^{-1} = \begin{pmatrix} \frac{2\mu_1 + M + 2}{12M(\mu_1 + M + 2)} & \frac{1}{4(\mu_1 + M + 2)} & \frac{2\mu_1 + M + 2}{4M\mu_1(\mu_1 + M + 2)} & \frac{1}{4\mu_1(\mu_1 + M + 2)} \\ \frac{2\mu_2 + M + 2}{12M(\mu_2 + M + 2)} & \frac{1}{4(\mu_2 + M + 2)} & \frac{2\mu_2 + M + 2}{4M\mu_2(\mu_2 + M + 2)} & \frac{1}{4\mu_2(\mu_2 + M + 2)} \\ 0 & 0 & \frac{1}{M+1} & -\frac{1}{M+1} \\ 0 & 0 & \frac{1}{M+1} & \frac{M}{M+1} \end{pmatrix},$$

and the equality

$$e^{tQ} = P \begin{bmatrix} e^{\mu_1 t} & 0 & 0 & 0 \\ 0 & e^{\mu_2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} P^{-1},$$

a proof of the following proposition is obtained.

Proposition 12. *The transition kernel $P_t = e^{tQ}$ is given by*

$$P_t = e^{\mu_1 t} A(\mu_1) + e^{\mu_2 t} A(\mu_2) + B,$$

where

$$A(\mu) = \frac{1}{\delta(M,\mu)} \begin{bmatrix} 2\mu + M + 2 & 3M & \frac{3(2\mu+M+2)}{\mu} & \frac{3M}{\mu} \\ M & 2\mu + 3M + 6 & \frac{3M}{\mu} & \frac{2\mu+3M+6}{\mu} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$B = \begin{bmatrix} 0 & 0 & b & 1-b \\ 0 & 0 & 1-b & b \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

with $\delta(M,\mu) = 4(\mu + M + 2)$ and $b = \dfrac{M+2}{2(M+1)}$.

The IICRs $\lambda_i(\cdot)/3$, for the initial sample configurations $i = 1$ and $i = 2$, correspond to the size-change functions

$$\lambda_i(t) = \frac{3(1 - P_t(i,3) - P_t(i,4))}{3P_t(i,1) + P_t(i,2)},$$

and the following proposition is verified.

Proposition 13. *When $t \to \infty$, $\lambda_i(\cdot)$, $i = 1, 2$, have the following limit*

$$\lim_{t\to\infty} \lambda_i(t) = -\frac{3}{\mu_1}.$$

In Figure 7 we plot the two functions $\lambda_i(\cdot)$, $i = 1, 2$ for $n = 2$ demes and $M = 1$ (left), respectively $M = 0.1$ (right); the dashed red line indicates the common asymptotic value $-\frac{3}{\mu_1}$.

Proof. Using Proposition 12 we get

$$\lim_{t\to\infty} \lambda_i(t) = -\frac{3(A(\mu_1)(i,3) + A(\mu_1)(i,4))}{3A(\mu_1)(i,1) + A(\mu_1)(i,2)}.$$

Then the relations $3A(\mu_1)(i,1) = \mu_1 A(\mu_1)(i,3)$ and $A(\mu_1)(i,2) = \mu_1 A(\mu_1)(i,4)$ allow to prove the result. □

The conditional cumulative distribution functions $\mathbb{P}(T_2^{(3),\lambda_i} \le \cdot | T_3^{(3),\lambda_i} = t)$ and $\mathbb{P}_i(T_2^{(3),2,M} \le \cdot | T_3^{(3),2,M} = t)$ are given, for every $t > 0$, by formulas analogous to (11) and (18):

$$\begin{aligned} \mathbb{P}_i(T_2^{(3),2,M} \le u | T_3^{(3),2,M} = t) &= \frac{F_{2,s}(u)\frac{d}{dt}P_t(i,3) + F_{2,d}(u)\frac{d}{dt}P_t(i,4)}{f_{T_3^{(3),2,M},i}(t)} \\ &= \frac{3F_{2,s}(u)P_t(i,1) + F_{2,d}(u)P_t(i,2)}{3P_t(i,1) + P_t(i,2)}, \end{aligned} \tag{27}$$

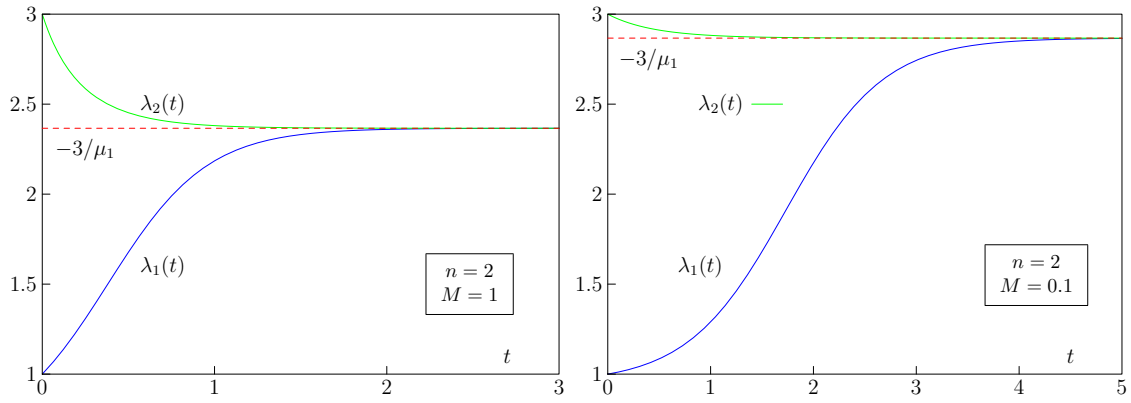

Figure 7: Plot of the functions $\lambda_i(\cdot),\ i = 1, 2$ for $n = 2$ and $M = 1$ (left), respectively $M = 0.1$ (right). The dashed red line corresponds to the asymptotic value $-3/\mu_1$.

$$\begin{aligned}\mathbb{P}(T_2^{(3),\lambda_i} \leq u | T_3^{(3),\lambda_i} = t) &= 1 - \exp\{-(\Lambda_i(t+u) - \Lambda_i(t))\} \\ &= 1 - \left(\frac{1 - P_{t+u}(i,3) - P_{t+u}(i,4)}{1 - P_t(i,3) - P_t(i,4)}\right)^{1/3}. \end{aligned} \tag{28}$$

In order to compare, for $u, t > 0$, the conditional cumulative distribution functions given in Equations (27) and (28), let us introduce the functions

$$g_i(u,t) := \mathbb{P}_i(T_2^{(3),2,M} \leq u | T_3^{(3),2,M} = t) - \mathbb{P}(T_2^{(3),\lambda_i} \leq u | T_3^{(3),\lambda_i} = t),\ i = 1, 2.$$

Proposition 14. *The functions $g_i(u,t),\ i = 1, 2$ have the following asymptotic behaviour :*

1. *For (u,t) in the neighborhood of $(0,0)$, we have*

$$\begin{aligned} g_1(u,t) &= \frac{M}{2}ut + o(u^2+t^2), \\ g_2(u,t) &= -\frac{u}{3} + \frac{6M+1}{18}u^2 + \frac{7M}{6}ut + o(u^2+t^2). \end{aligned}$$

2. *For fixed $t > 0$, when $u \to +\infty$,*

$$g_i(u,t) = -K_{1,i}(M,t)e^{-\beta u} + o(e^{-\beta u}),\ i = 1, 2,$$

where $K_{1,i}(M,t) > 0$ is given by

$$K_{1,i}(M,t) = \frac{3P_t(i,1)}{3P_t(i,1) + P_t(i,2)}\frac{1-a}{\beta} - \frac{P_t(i,2)}{3P_t(i,1) + P_t(i,2)}\frac{c}{\beta},$$

where constants β, a, c are defined in equations (9) and (10) with $n = 2$, i.e.

$$\beta = M + \frac{1}{2}(1 - (4M^2+1)^{1/2}),\ a = \frac{1}{2}(1 + (4M^2+1)^{-1/2}),\ c = -M(4M^2+1)^{-1/2}.$$

3. *For fixed $u \geq 0$, we have*

$$\lim_{t \to +\infty} g_i(u,t) = -K_3(M,u),\ i = 1, 2,$$

where $K_3(M,u) > 0$ *is given by*

$$K_3(M,u) = c_1 e^{-\beta u} + c_2 e^{-\alpha u} - e^{\frac{\mu_1}{3}u},$$

with

$$\alpha = M + \frac{1}{2}(1 + (4M^2+1)^{1/2}),\ \beta = M + \frac{1}{2}(1 - (4M^2+1)^{1/2}),$$
$$\phi(\mu_1) = \frac{3M}{2(\mu_1 + 3M + 3)},\ c_1 = \frac{\phi(\mu_1) - \alpha}{\beta - \alpha},\ c_2 = \frac{\beta - \phi(\mu_1)}{\beta - \alpha}.$$

The proof is similar to the one given in the case $n > 2$.

Most of the calculations of appendices A and B were made and/or verified using the computer algebra system Maple: programs and tracks of their execution are available from the authors.

Aknowledgements: The authors wish to thank Josué M. Corujo Rodriguez for very interesting discussions in preparing this article. This research was funded through the 2015-2016 BiodivERsA COFUND call for research proposals, with the national funders ANR (ANR-16-EBI3-0014), FCT (Biodiversa/0003/2015) and PT-DLR (01LC1617A), under the INFRAGECO (Inference, Fragmentation, Genomics, and Conservation) Project (https://infrageco-biodiversa.org/). The research was also supported by the LABEX entitled TULIP (ANR-10-LABX-41), as well as the Pôle de Recherche et d'Enseignement Suprieur (PRES) and the Région Midi-Pyrénées, France. We finally thank the LIA BEEG-B (Laboratoire International Associ - Bioinformatics, Ecology, Evolution, Genomics and Behaviour) (CNRS) and the PESSOA program for facilitating travel and collaboration between EDB, IMT and INSA in Toulouse and the IGC, in Portugal.

The authors declare no conflict of interest.

Bibliography

References

Mark A. Beaumont. Detecting population expansion and decline using microsatellites. *Genetics*, 153(4):2013–2029, 1999.

Mark A. Beaumont. Recent developments in genetic data analysis: what can they tell us about human demographic history? *Heredity*, 92(5):365–379, 2004.

Lounès Chikhi, Vitor C. Sousa, Pierre Luisi, Benoit Goossens, and Mark A. Beaumont. The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics*, 186(3):983–995, 2010. doi: 10.1534/genetics.110.118661.

Lounès Chikhi, Willy Rodriguez, Simona Grusea, Patricia Santos, Simon Boitard, and Olivier Mazet. The IICR (inverse instantaneous coalescence rate) as a summary of genomic diversity: insights into demographic inference and model choice. *Heredity*, 2018.

Benoit Goossens, Lounès Chikhi, Marc Ancrenaz, Isabelle Lackman-Ancrenaz, Patrick Andau, Michael W. Bruford, et al. Genetic signature of anthropogenic population collapse in orang-utans. *PLoS Biology*, 4(2):285, 2006.

RC Griffiths and Simon Tavaré. The age of a mutation in a general coalescent tree. *Stochastic Models*, 14(1-2):273–295, 1998.

Robert C. Griffiths and Simon Tavaré. Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 344(1310):403–410, 1994. ISSN 0962-8436. doi: 10.1098/rstb.1994.0079.

Rasmus Heller, Lounes Chikhi, and Hans Redlef Siegismund. The confounding effect of population structure on bayesian skyline plot inferences of demographic history. *PLoS ONE*, 8 (5):e62992, 2013. doi: 10.1371/journal.pone.0062992.

Hilde M. Herbots. *Stochastic models in population genetics: genealogy and genetic differentiation in structured populations*. PhD thesis, University of London, 1994. URL `https://qmro.qmul.ac.uk/xmlui/handle/123456789/1482?show=full`.

Richard R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2):183 – 201, 1983. ISSN 0040-5809. doi: http://dx.doi.org/10.1016/0040-5809(83)90013-8.

Richard R. Hudson. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002. doi: 10.1093/bioinformatics/18.2.337.

Richard R. Hudson et al. Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology*, 7(1):44, 1990.

John F.C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235 – 248, 1982. doi: http://dx.doi.org/10.1016/0304-4149(82)90011-4.

Serge Lang. *Algebra*. Springer, rev. 3rd edition, 2002.

Olivier Mazet, Willy Rodriguez, and Lounès Chikhi. Demographic inference using genetic data from a single individual: Separating population size variation from population structure. *Theoretical Population Biology*, 104:46 – 58, 2015. doi: http://dx.doi.org/10.1016/j.tpb.2015.06.003.

Olivier Mazet, Willy Rodriguez, Simona Grusea, Simon Boitard, and Lounès Chikhi. On the importance of being structured: instantaneous coalescence rates and human evolution–lessons for ancestral population size inference? *Heredity*, 116(4):362–371, 2016. ISSN 1365-2540. doi: 10.1038/hdy.2015.104.

M Nei and N Takahata. Effective population size, genetic diversity, and coalescence time in subdivided populations. *J. Mol. Evol.*, 37(3):240–4, 1993.

Richard W.D. Nickalls. A new approach to solving the cubic: Cardan's solution revealed. *The Mathematical Gazette*, 77:354–359, 1993. URL `http://www.nickalls.org/dick/papers/maths/cubic1993.pdf`.

Rasmus Nielsen and John Wakeley. Distinguishing migration from isolation: A Markov Chain Monte Carlo approach. *Genetics*, 158(2):885–896, 2001.

James R. Norris. *Markov Chains.* Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, 1998. ISBN 978-0-521-48181-6.

M. Notohara. The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology*, 29(1):59–75, 1990. ISSN 0303-6812. doi: 10.1007/BF00173909. URL `http://dx.doi.org/10.1007/BF00173909`.

Benjamin M. Peter, Daniel Wegmann, and Laurent Excoffier. Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. *Molecular Ecology*, 19(21):4648–4660, 2010.

Erwan Quéméré, Xavier Amelot, Julie Pierson, Brigitte Crouau-Roy, and Lounès Chikhi. Genetic data suggest a natural prehuman origin of open habitats in northern Madagascar and question the deforestation narrative in this region. *Proceedings of the National Academy of Sciences*, 109(32):13028–13033, 2012.

Willy Rodriguez. *Estimation de l'histoire démographique des populations à partir de génomes entièrement séquencés.* PhD thesis, University Paul Sabatier, Toulouse, France, June 2016.

Willy Rodriguez, Olivier Mazet, Simona Grusea, Didier Pinchon, Simon Boitard, and Lounes Chikhi. The IICR, Q-matrices, and the structured coalescent: towards a general framework of demographic inference with arbitrary changes in population structure, connectivity and size. *In preparation*, 201X.

Alan R. Rogers and Henry Harpending. Population growth makes waves in the distribution of pairwise genetic differences. *Molecular biology and evolution*, 9(3):552–569, 1992.

Stephan Schiffels and Richard Durbin. Inferring human population size and separation history from multiple genome sequences. *Nat Genet*, 46(8):919–925, August 2014. ISSN 1061-4036. URL `http://dx.doi.org/10.1038/ng.3015`.

Montgomery Slatkin. Inbreeding coefficients and coalescence times. *Genetical Research*, 58(02): 167–175, 1991.

Jay F Storz and Mark A Beaumont. Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical bayesian model. *Evolution*, 56(1):154–166, 2002.

Fumio Tajima. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105:437–460, 1983.

S.S. Vallander. Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability and its Applications*, 18:784–786, 1973.

John Wakeley. Nonequilibrium migration in human history. *Genetics*, 153(4):1863–1871, 1999.

Daniel B Weissman and Oskar Hallatschek. Minimal-assumption inference from population-genomic data. *eLife*, 6, 2017.

Hilde M. Wilkinson-Herbots. Genealogy and subpopulation differentiation under various models of population structure. *Journal of Mathematical Biology*, 37(6):535–585, 1998. ISSN 0303-6812. doi: 10.1007/s002850050140.

Sewall Wright. Evolution in mendelian populations. *Genetics*, 16(2):97, 1931. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1201091/`.