Journal de la Société Française de Statistique
Vol. 159 No. 3 (2018)

Simulation of stochastic models of structured population in population genetics under neutrality

Titre: Simulation de modèles stochastiques de populations structurées en génétique des populations sous neutralité

Pierre Pudlo[1] and Mohammed Sedki[2]

Abstract: This paper describes some population genetic models under neutrality, involving genetic drift and mutations. Starting with Kingman's coalescent we show how structured populations can be modeled. We detail these models by showing how simulation algorithms can be written. In particular we highlight the latent processes than rule out the explicit computation of the likelihood function on a dataset.

Résumé : Cet article décrit quelques modèles de génétique des populations sous neutralité, incluant dérive génétique et mutations. À partir du coalescent de Kingman, nous montrons comment on peut modéliser des populations structurées. Nous détaillons ces modèles en montrant comment il est possible d'écrire des algorithmes de simulations. En particulier, nous mettons en avant l'ensemble des processus latents qui rendent le calcul de la fonction de vraisemblance sur un jeu de données difficile, voire impossible.

Keywords: population genetics, Kingman coalescent process, intractable likelihood
Mots-clés : génétique des populations, processus de coalescence de Kingman, vraisemblance intraitable
AMS 2000 subject classifications: 60-08, 92D15, 92D40, 62P10, 65C60

1. Introduction

Population genetics concerns the distribution of genetic polymorphism within populations of individuals of the same species. Two mechanisms govern the distribution of the various genetic states, named alleles, and their evolution over time: a mutational process (mainly a Markov process on the state space of all possible alleles) and variations of the frequencies of alleles from one generation to another because of the varying number of children per individual. The latter process, named genetic drift, is governed by the its structuration into sub-populations and the size of each one. Genetic neutrality (Kimura, 1968, 1983) assumes that the various allelic states at a given position of the genome (named locus) are neither beneficial nor detrimental to the individuals that carry them. The neutral hypothesis is certainly not realistic accros the whole genome. It is well known that certain loci undergo selective pressure, *i.e.* some allelic states at this position of the genome can confer benefits or disadvantages to their carriers, whose reproductive success is greatly influenced by them. However, under neutrality, the number of children of a given individual is independent of their genetic type and the two evolutionary processes described

[1] Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France.
E-mail: pierre.pudlo@univ-amu.fr
[2] Université Paris-Sud, unité Inserm 1018. Centre d'épidémiologie et Santé des Populations, Paris, France.
E-mail: mohammed.sedki@u-psud.fr

 ISSN: 2102-6238

above (mutation and genetic drift) are independent. The statistical challenge is therefore to infer the history of populations from polymorphism data taken from a current sample of individuals. The present paper does not present inferential methods to answer this question. Rather, the aim is to describe the stochastic processes that allow demographic and mutational parameters to be linked to genetic data. In other words, our objective is to explain the models that allow a proper likelihood to be defined.

We discuss a large family of stochastic models under neutrality that are part of the scientific folklore of population genetics, but which are rarely described with accuracy. The simplest way to present these models is to provide simulation algorithms. Writing mathematical formulas from these algorithms is left to the reader, if need be. Furthermore, we aim to demonstrate that, given genetic data, a likelihood is not an explicit function of the parameters, but represents as an integral over a large latent process that encompasses the past histories of the ancestors of the individuals comprising the sample.

For each individual in the sample, the genetic information that we consider in the data $\mathbf{x}$ is limited. We are interested only by a few given positions in the genome called loci. At these loci, the DNA sequence varies from one individual to the other due to genetic polymorphism, *i.e.* the mutations that occurred during the evolution of the species. Variants are called alleles or allelic states. We denote by $\mathbf{x}_j$ the genetic data at locus j and by L the number of loci so that $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_L)$.

Let us denote by ϕ the set of parameters of our model. In the present paper, we consider only loci from the nuclear genome and assume that they are independent. Thus our likelihood $f(\mathbf{x}|\phi)$ is a product of likelihoods evaluated on single locus dataset:

$$f(\mathbf{x}|\phi) = \prod_{j=1}^{L} f(\mathbf{x}_j|\phi). \tag{1}$$

This assumption actually hides a composite likelihood model Hudson (2001); Stumpf and McVean (2003). Owing to genetic mixing due to recombination, this approximation is valid if the loci are sufficiently distant from each other, so as to be able to consider the different genealogies of the loci as independent. It should be pointed out here that more complex models have been proposed to take ancestral recombination into account Griffiths and Marjoram (1997). Note also that the set of parameters ϕ can be broken down into two subsets, one ϕ_{dem} concerning the demography of the species, and the other ϕ_{mut} the mutational processes.

A dataset is composed of individuals sampled from different populations of interest (sometimes named colonies or demes). We have numbered the populations with integer numbers from 1 to D, and labelled them $Pop1$ to $PopD$. Unlike many statistical problems, the memberships of individuals are known. The typical sample size of a sample varies from about 20 to one hundred per population. The sample size of individuals arising from $Popi$ is denoted n_i. The random models that we present here explain genetic polymorphism by tracing the evolution of a species and its mutations over time. We must thus specify which forms of genetic information can be found at each locus. In fact, there are three types of locus: microsatellite, sequence and single nucleotide polymorphism(SNP). We focus here on microsatellite loci which are DNA sequences where a short motif (typically of 1 to 4 base pairs) is repeated. Because of these repeats whose number varies between individuals, the microsatellites are sources of polymorphism. We describe two classical, mutational models on this type of locus in Section 3.

 ISSN: 2102-6238

The genetic data $\mathbf{x}_j$ at locus j can be explained by some biological concepts and evolutionary phenomena. A demographic scenario is a series of spatio-temporal events sorted from the most recent to the oldest one (Figure 1). Stochastic models explain the genetic data with a latent process that is constrained by the demographic scenario. The latent part of the model includes a genealogy $\mathscr{G}_j$ and a set of ancestral genotypes $\mathscr{M}_j$. So that the likelihood of the genetic data $\mathbf{x}_j$ at locus j can be written as

$$f(\mathbf{x}_j|\phi) = \iint_{\{M_j \to x_j\}} h(\mathscr{M}_j|\mathscr{G}_j, \phi_{\text{mut}}) g(\mathscr{G}_j|\phi_{\text{dem}}) \, \mathrm{d}\mathscr{M}_j \, \mathrm{d}\mathscr{G}_j, \tag{2}$$

where

— $g(\mathscr{G}|\phi_{\text{dem}})$ is the density associated to the distribution of the genealogy $\mathscr{G}$ with respect to some reference measure $\mathrm{d}\mathscr{G}$,

— $h(\mathscr{M}|\phi_{\text{mut}})$ is the density associated to the distribution of the mutational process $\mathscr{M}$ with respect to another reference measure $\mathrm{d}\mathscr{M}$, knowing the genealogy $\mathscr{G}$,

— $\{\mathscr{M}_j \to \mathbf{x}_j\}$ is the set of paths of the mutational process which leads to labels on the tips of the genealogy that are equal to the observed dataset $\mathbf{x}_j$

A genealogy $\mathscr{G}_j$ can be visualized with a dendrogram. A lineage of the dendrogram corresponds to the ancestry of an individual until the most recent common ancestor (MRCA) of the observed sample is reached. Simulations of genealogies $\mathscr{G}_j$ according to the distribution $g(\mathscr{G}_j|\phi_{\text{dem}})$ are given in Section 2. The evolution of the lineages of a genealogy is governed by various in-between population events in the demographic scenario (Section 2) whose parameters are in the set ϕ_{dem}. Given a mutation model and its parameters that lie in the set ϕ_{mut}, the simulation of genotypes $\mathscr{M}_j$ of the ancestral individuals according to the distribution $h(\mathscr{M}_j|\mathscr{G}_j, \phi_{\text{mut}})$, from the MRCA to the tips of the genealogy $\mathscr{G}_j$, is described in Section 3. According to the kind of genetic data, a mutational model (Section 3) generates the MRCA genotype and mutates its state along the lineages at the times set by the point process.

The latent part $(\mathscr{G}_j, \mathscr{M}_j)$ of the process adds a temporal dimension to the stochastic models that is not present in the data. Thus, the understanding of the parameters of these models depends on a time scale that is difficult to set (Section 4). The computation of the likelihood $f\left(\mathbf{x}\middle|\phi\right)$ requires a marginalization that can be achieved by integrating over the latent part (Section 4).

2. Sample genealogies

Some stochastic models (Wright-Fisher, Moran, etc. see Wakeley, 2005, Chapter 3) explain the evolution of the whole population from past to present, and then sample the individuals from the last generation. When the population size is large, simulating data according to these models can be very slow. The method presented below simulates only the history of the sampled individuals.

Section 2 describes the simulation of a genealogy $\mathscr{G}_j$ according to $g(\mathscr{G}_j|\phi_{\text{dem}})$. Note the simulation starts from the present, since the time at which the MRCA is met is random and unknown. First we deal with a very simple demographic scenario of a single closed population at equilibrium. We introduce here the fundamental tool to simulate genealogies which is the coalescent process of Kingman (1982a,c,b). A complete description of the Kingman coalescent process can

 ISSN: 2102-6238

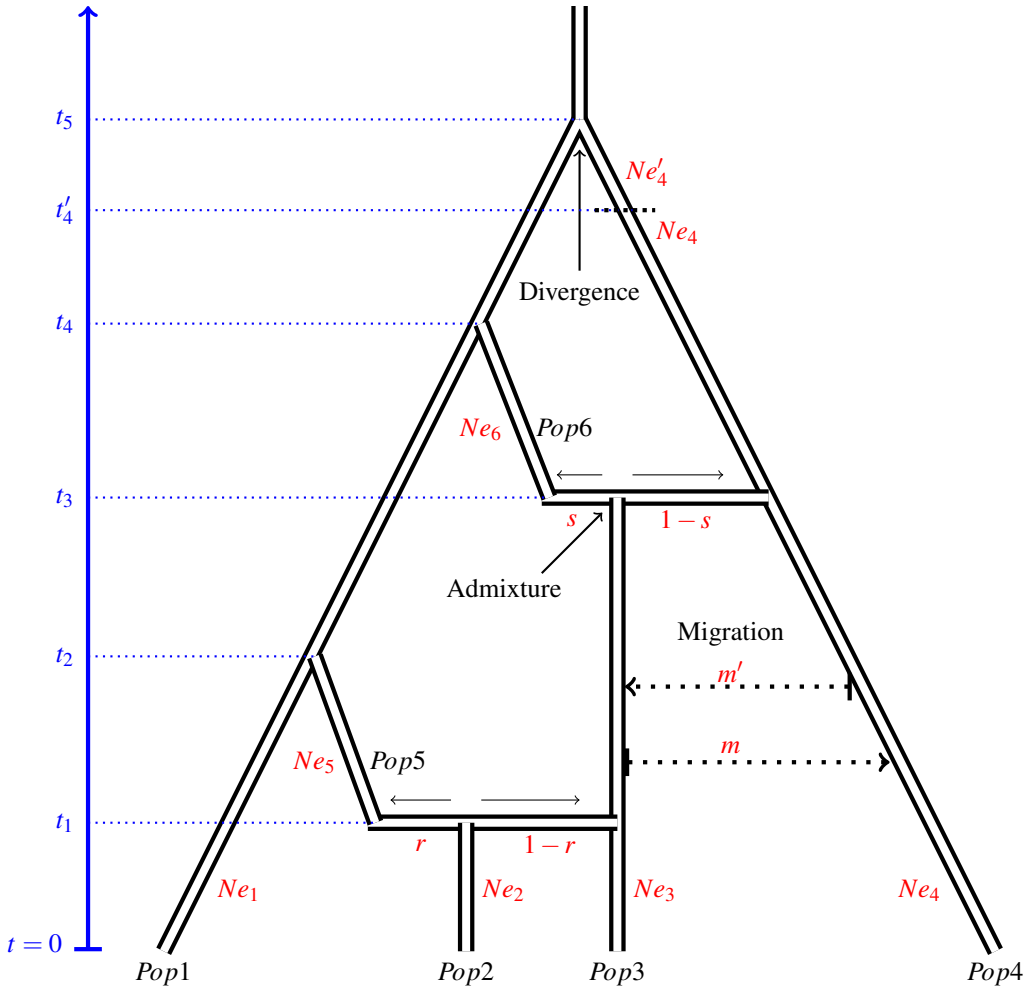

FIGURE 1. *Example of a complex evolutive scenario composed of in-between population events. The scenario deals with four sampled s* $Pop1,\ldots,Pop4$ *et two other non-observed populations* $Pop5$ *and* $Pop6$*. The branches of the above picture are "tubes" and the demographic scenario constraints the genealogy to stay inside those "tubes". Migration between populations* $Pop3$ *et* $Pop4$ *during the period* $[0,t_3]$ *is parametrized by two migration rates* m *and* m'*. The two admixture events are parametrized by dates* t_1 *and* t_3*, as well their respective admixtures rates* r *and* s*. The three other events are divergences, respectively at times* t_2*,* t_4 *and* t_5*. The event at time* t'_4 *corresponds to a change in the effective population size of population* $Pop4$*.*

 ISSN: 2102-6238

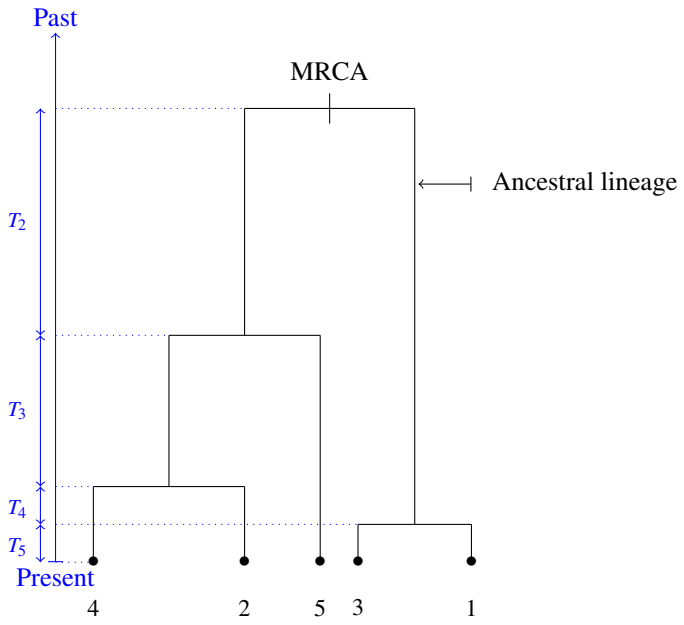

FIGURE 2. *Example of a genealogy of $n = 5$ individuals from a closed population at equilibrium. The sampled individuals stand on the tips of the dendrogram; the coalescent time $T_2, \ldots, T_5$ are independent, and T_k is drawn from an exponential distribution of rate $k(k-1)/2$.*

be found in Sainudiin et al. (2015). Then, we deal with the genealogy of structured populations. These populations are structured by interpopulation events. We describe the evolution of the lineages of a genealogy according to these events such as divergence, admixture and migration.

2.1. A single closed population: the Kingman Coalescent

The Kingman coalescent process (Kingman, 1982a,c,b) is the fundamental tool to simulate genealogies. For the sake of clarity, we focus here on the simplest scenario that considers only a single closed population at equilibrium. We assume that this population is neither subjected to an external flow of genes, nor undergoes any internal demographic variation.

The gene genealogy of the sample is represented by a dendrogram (Figure 2). We generate ancestral lineages until we reach the most recent common ancestor (MRCA). A coalescent event occurs when the lineages of two individuals join at a node of the dendrogram (Figure 2). The genealogy of a sample of k individuals is thus composed of $k-1$ coalescent events. Each coalescent event decreases the number of ancestral lineages by 1 until we reach the last lineage at the root of the dendrogram, which corresponds to the MRCA.

The variables denoted T_i represent the time between successive coalescent events (Figure 2). The distribution of the genealogy of k individuals is fully characterized by the distribution of the draw of the two lineages that coalesce at the coalescent event and the time between these events $T_k, \ldots, T_2$. According to the Kingman coalescent process, the times between coalescent events are independent and T_k is distributed according to the exponential distribution of the parameter (or rate) $k(k-1)/2$.

 ISSN: 2102-6238

Algorithm 1 Kingman coalescent in natural time

INPUT: The sample size n, and the effective population size Ne.

Set $k = n$.

While $k \geq 2$ **do**

1) Draw the inter-coalescent time T_k from the exponential distribution of parameter $k(k-1)\Big/(2Ne)$.

2) Add T_k to the length of the k lineages.

3) Among the k lineages, pick two lineages randomly and join them to make a new node of the dendrogram.

4) $k \leftarrow k-1$.

EndWhile

We can describe a coalescent event as follows when the number of ancestral lineages is k. For each of the $\binom{k}{2}$ pairs of lineages in competition for coalescing, we start a random clock of exponential distribution with rate 1. The clock that rings first, namely the minimal value, picks the pair of lineages that will coalesce. In practice, the draws of $\binom{k}{2}$ exponential distributions are slow. We can decrease the number of random variables we have to simulate with the following Lemma.

Lemme 1. *Let $T_1, \ldots, T_\ell$ be independent, exponentially distributed random variables of respective rates $\lambda_1, \ldots, \lambda_\ell$. Then, the random variable $T = \inf_{1 \leq i \leq \ell} T_i$ is distributed according to the exponential distribution of rate $\sum_{i=1}^{\ell} \lambda_i$. And independently, the random variable T is equal to T_k with probability $\lambda_k \Big/ \sum_{i=1}^{\ell} \lambda_i$.*

Hence, with Lemma 1, the simulation of a coalescent event boils down to drawing a variable from to the exponential distribution with rate $\binom{k}{2} = k(k-1)/2$ and to drawing a pair of lineages at random among the $\binom{k}{2}$ lineages.

The time scale of the above mentioned description is such that a period of time of 1 unit corresponds to Ne generations, where Ne is a parameter of the model named the effective population size of the population (more details in Section 4). Algorithm 1 describes the Kingman coalescent process of a sample of k individuals taken from a population of constant effective population size Ne. Hence, in this algorithm, the time scale is the "natural time scale", and the coalescent rate is multiplied by Ne.

Finally, note that the order of the individuals at the bottom of the dendrogram can differ from the order of the individuals in the sample so that the lineages do not coalesce. In the genealogy given in Figure 2, the numbers at the bottom of the dendrogram are the numbers of the individuals in the dataset.

In this simple case, we can write the distribution of the genealogy explicitly. To this end, we encode the genealogy $\mathscr{G}_j$ by a sequence of triplets (x_k, y_k, T_k) where

— x_k, y_k denotes the pair of lineages that has coalesced when there are k lineages

— and T_k is the in-between-coalescence time.

For example, the genealogy of Figure 2 is encoded by $\{(4,5,T_5),(1,2,T_4),(1,2,T_3),(1,2,T_2)\}$.

 ISSN: 2102-6238

Then, the distribution of $\mathscr{G}_i$ is given by

$$g(\mathscr{G}_j|Ne) \propto \prod_{k=n,n-1,\ldots,2} \frac{k(k-1)}{2Ne} \exp\left(-\frac{k(k-1)}{2Ne} T_k\right). \tag{3}$$

2.2. *Many structured populations*

Algorithm 2 Partial genealogy in an independent population

INPUT: The ancestral sample size n of the population i at time t, the effective population size Ne and the dates t, t'.

Set $k = n$ and simulate T_k from an exponential distribution with parameter $k\big(k-1\big)/2Ne$.
While $\big(t + T_k\big) \leq t'$ **do**
1) Increase the lengths of all the k lineages by T_k.
2) Draw independently among the k lineages a pair of lineages that is gathered to form a node of the dendrogram at time $t + T_k$.
3) Set $k \leftarrow k-1$, $t \leftarrow t + T_k$
4) Draw a time T_k from the exponential distribution with rate $k\big(k-1\big)/2Ne$.
EndWhile
Set $n' = k$
If $\big(t + T_k\big) > t'$ **Then**
Increase the lengths of all lineages time t' is reached.
EndIf

Algorithm 3 Genealogy of populations in presence of migration, in coalescent time

INPUT: The sizes and the ancestral samples: $k_1, \ldots, k_D$ and the migration rates m_{ij}.

For i =1 $\rightarrow D$ **do**
1) Associate an exponential random clock with rate $1/Ne_i$ to each pair of ancestral individuals in population i(which corresponds to a potential coalescent event)
2) Associate $D-1$ exponential random clocks of parameters $m_{ij}, 1 \leq j \neq i \leq D$ to each ancestral individual of population i (which correspond to potential migration events).
EndFor
Among all these clocks in competition, the first that rings, *i.e.* the smallest one, wins. If the winning clock corresponds to a pair of individuals, a coalescent event occurs: we add a node to the genealogy joining the attached pair at the time the clock rings. If the winning clock is attached to a single individual, and has parameter m_{ij}, then the ancestral lineage of this individual is sent from population i to population j.

We now describe the distribution of a genealogy following an demographic scenario whose structure is governed by in-between population events. We combine these events with Kingman coalescent process which describes the within population evolution. We begin by describing the three kinds of inbetween population events.

— Divergence (Figure 3, (a)) is the coalescing of two populations (backward in time) at a given date.

 ISSN: 2102-6238

Algorithm 4 Genealogy of 2 populations in presence of migration

INPUT: The size of the ancestral samples k_1 and k_2, the migration rates m_{12} and m_{21}, and the effective sample sizes Ne_1 and Ne_2 of the populations.

1) Choose population numbered i among the two populations with probability

$$\frac{k_i m_{ij} + \left[k_i(k_i-1)/2Ne_i\right]}{k_1 m_{12} + \left[k_1(k_1-1)/2Ne_1\right] + k_2 m_{21} + \left[k_2(k_2-1)/2Ne_2\right]}.$$

2) Choose the type of event that will occur: either a coalescent event with probability

$$\frac{k_i(k_i-1)/2Ne_i}{k_i m_{ij} + \left[k_i(k_i-1)/2Ne_i\right]},$$

or a migration event with probability

$$\frac{k_i m_{ij}}{k_i m_{ij} + \left[k_i(k_i-1)/2Ne_i\right]}.$$

3) **If** the event is a coalescent event within population i:
- Simulate T_c from an exponential distribution with rate $k_i(k_i-1)/2Ne_i$.
- Increase the lengths of all lineages by T_c.
- Draw uniformly at random the pair of lineages that coalesces between the k_i lineages of population i and draw the corresponding node in the dendrogram.
- Set $k_i \leftarrow k_i - 1$, and go back to **1)**.

EndIf

4) **If** the event is a migration event from population i to population j:
- Simulate T_m from an exponential distribution with rate $k_i m_{ij}/Ne_i$.
- Increase the lengths of all lineages by T_m.
- Migrate a lineage drawn uniformly at random within population i to send it to population $j \neq i$.
- Set $k_i \leftarrow k_i - 1$ and $k_j \leftarrow k_j + 1$, and go back to **1)**.

EndIf

— Admixture (Figure 3, (b)) is the splitting of one population into two parts at the time of the event. The lineages are sent randomly to both populations, with a given probability of going to the first one, often named the admixture rate.

— Migration (Figure 3, (c)) allows lineages to move from one population to another over a period of time, according to rates expressed by time unit and per gene.

Divergence and admixture are both instantaneous events while migration is an event that lasts over a given period of time.

The algorithm that produces the genealogy in these three cases simulates the evolutions of the lineages between these events and the instantaneous changes due to these evolutions. In the presence of a migration event over a period of time, the evolution of the lineages of the genealogy is somewhat different. Coalescent and migration events compete over the period. Algorithm 5 sums up the simulation steps of a genealogy constrained to follow a given demographic scenario.

Genealogy between two instantaneous events The genealogy within a population between two dates (t and t', where $t' > t$, in Figure 3 (a) and (b)) of successive population events is

 ISSN: 2102-6238

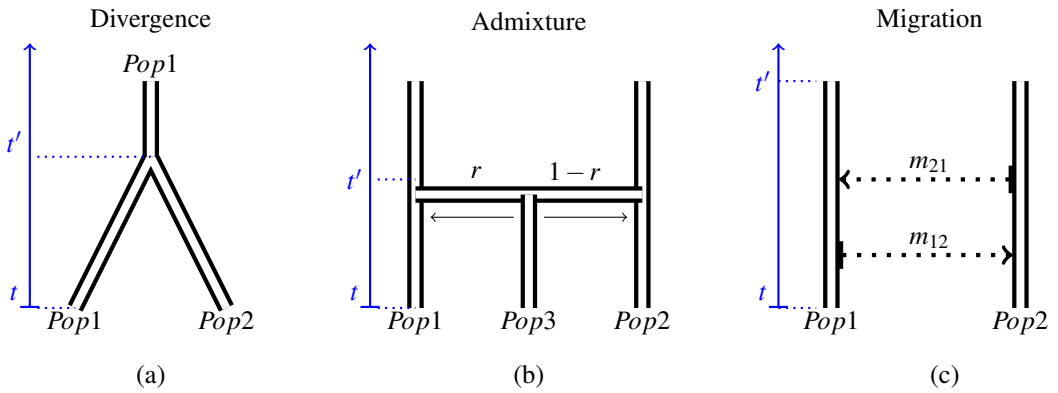

FIGURE 3. *Graphical representation of three kind of in-between population events. There exists two families of events between populations. The first family is simple and is made of instantaneous events. This is the case of divergence or admixture.* ***(a)*** *Two populations evolve and are joined in the case of a divergence.* ***(b)*** *Trois populations evolve in parallel in the case of an admixture. In this situation, each tube represents the genealogy (one can imagine that the genealogy is constrained to stay inside the tubes) of the population which evolves independently of the other ones according Kingman process.*
*The second family of event is reduced here to migration.****(c)*** *This situation is slightly more complex that the other ones because of gene flows. Here, a single process governs the evolution of the lineages of populations Pop*1 *et Pop*2*. Displacement of lineages from one population to the other one is in competition with coalescent events within each population.*

distributed according to the Kingman coalescent process independently of the other populations. Algorithm 2 describes the process. It differs slightly from Algorithm 1, which describes the evolution of a closed population at equilibrium until MRCA is reached. Indeed, Algorithm 2 describes the Kingman coalescent process over a period of time of length $t' - t$ between two population events. If the effective size of the population (Figure 1) changes at a given time $t'' \in [t; t']$, the period of time $[t; t']$ should be split into two parts, $[t; t'']$ and $[t''; t']$.

For instance, the contribution of a part drawn from Algorithm 2 to the distribution $g(\mathscr{G}_j|\phi_{\text{dem}})$ between times t and t' is given by

$$\sum_{n'=1}^{n} \prod_{k=(n'+1)}^{n} \frac{k(k-1)}{2Ne} \exp\left(\frac{k(k-1)}{2Ne} T_k\right) \times \exp\left(\frac{n'(n'-1)}{2Ne}(t' - t - T_n - \cdots - T_{n'+1})\right) \quad (4)$$

with the notations set in the Algorithm. Note that the last term of this product is the probability that $t + T_n + T_{n-1} + \cdots + T_{n'} > t'$, and that, if $n' = 1$, the last term is equal to 1. Note also that n' is the number of ancestral lineages remaining in the population just before time t'. Hence other contributions to the distribution of $\mathscr{G}_j$ that depends on this size should be added inside the sum over possible values of n'.

Divergence At the time of the divergence event, the lineages that remain in both populations (*Pop*1 and *Pop*2 of Figure 3 (a)) are gathered in a single population (*Pop*1 of Figure 3 (a)). Such events are deterministic and do not add any term to distribution of $g(\mathscr{G}_j|\phi_{\text{dem}})$.

 ISSN: 2102-6238

Admixture At the time of the admixture event, the ancestral sample of $Pop3$ (see Figure 3, (b)) is split into two other populations as follows: a lineage of population $Pop3$ is sent to $Pop1$ with probability r and to $Pop2$ with probability $(1-r)$, where r is a parameter of the model named the admixture rate. The contribution of such an event to the distribution $g(\mathscr{G}_j|\phi_{\text{dem}})$ is that of a binomial distribution with parameter r.

Migration In Algorithm 3, we describe the evolution of ancestral lineages when migrations occur between D populations. The migration is parametrized by the migration rates from population i to population j, denoted m_{ij}. Algorithm 3 details all exponential clocks in competition and their own parameters. The clock that reaches the minimum time (i.e. that rings first) chooses the kind of event that occurs. In practice, the number of exponential clocks to be simulated is much smaller. As in the Section 2.1, we can simplify the Algorithm by using Lemma 1. Algorithm 4 is an example with $D = 2$ populations, and we refer the reader to (Wakeley, 2005, Chapter 5) for explicit details on the case $D > 2$. The genealogy described in this algorithm corresponds to the scenario in Figure 3 (c). It is used on populations $Pop1$ and $Pop2$ until time t' is reached. Alas, we are unable to describe the contribution of migration events to the distribution $g(\mathscr{G}_j|\phi_{\text{dem}})$ as in (3) or (4), since it is much more difficult in the cases we consider here.

Finally, we presente the general scheme to simulate a genealogy constrained by the events of a demographic scenario in Algorithm 5. The Kingman coalescent process indeed constitutes the cornerstone of these algorithms to simulate gene genealogies. It is used piecewise, or in competition with an independent migration process. From the evolution of populations of interest, these algorithms make it possible to extract the only part of the process that matters regarding the sample of individuals in the dataset. Note that some approximations of the migration process are proposed to avoid D populations being considered in parallel, see e.g. Müller et al. (2017) and the references therein.

Algorithm 5 Genealogy constrained by a demographic scenario

Sort the instantaneous in-between population events from the most recent to the oldest one.

For t going from the most recent event to the oldest one **do**

1) Simulate the genealogies within each population: the Kingman coalescent process is drawn independently for each population until time t, or processes that combine Kingman coalescent processes and migration processes are drawn.

2) Apply the instantaneous in-between population event at time t.

EndFor

Simulate a single Kingman coalescent process or a combined migration-coalescent process until MRCA is reached.

3. Mutational processes

Let us now look at the distribution of the genotypes of the sample conditionally upon the genealogy. The simulation of the genealogy starts from $t = 0$ to the old age of the MRCA. Indeed, the genealogy traces the ancestors that gave their genes to the sample in the past of the populations.

 ISSN: 2102-6238

However, the simulation of the mutational process along the branches of the genealogy starts from the old age of the MRCA and progresses towards the present, since it models the mutations of the ancestors of the sample.

3.1. Mutations on microsatellite loci

We describe the positions of mutations with Poisson point process on the branches of the dendrogram. Then we introduce two mutational models for microsatellite loci. A Markov chain associated the mutational model applies mutations to their positions on the genealogy.

Positioning the mutations on the branches of the genealogy The mutation rate per unit of natural time and per individual diploid is given by the parameter μ. Conditional upon the genealogy, mutations occur on the dendrogram according to a Poisson point process with rate $\mu/2$. On a branch of length t, the distribution of the number N of mutations is Poisson with parameter $\mu t/2$, and the N random mutations are uniformly spread over the branch.

Algorithm 6 Mutational process on genealogy

INPUT: A genealogy $\mathscr{G}$, a mutation rate μ, and the transition matrix Q of the Markov chain with stationary distribution ν.

1) Apply the Poisson point process on the branches of the genealogy $\mathscr{G}$.
2) Sort the point mutations on $\mathscr{G}$ from the oldest to the newest.
3) Simulate the genotype of the MRCA from the distribution ν.
4) Sweep $\mathscr{G}$ from the MRCA to leaves : construct the genotypes along the branches.
— **if** a coalescent event is met, **then** duplicate the genotype above the node of the dendrogram.
— **if** a mutation event is met, **then** apply a one-step transition Markov chain Q on the branch (*i.e* the individual with the mutation).

Mutation models on microsatellite loci We introduce here two mutational models (Whittaker et al., 2003; Cornuet et al., 2006): the Stepwise Mutation Model (SMM) and the Generalized Mutation Model (GMM) which were specially designed for microsatellite loci . Both mutational models use a simple parameterization. Markov chains associated with GSM and SMM are symmetrical random walks on an interval of integer numbers, $[\![a;b]\!]$ of $\mathbb{N}$. Applying a one-step transition of the GSM Markov chain is equivalent to adding $\pm mG$ to the length of the locus, where m is the length of the repeated pattern (known), G is a geometric random variable with parameter p and $\pm$ is a random sign. In practice, the parameter $p \approx 0.2$. Whereas for the SMM model, a mutation is equivalent to adding the length of the locus by $\pm m$ base pairs, where m is the length of the repeated pattern, and $\pm$ is a random sign. Sometimes, applying mutations makes the genotype exceed the bounds a and b of all the alleles. In this case, we set the genotype to the closest value among the two bounds a, b of the state space. To simulate the genotypes of the sample at a given locus, we start by simulating the genotype of the MRCA, and then we let it evolve along the genealogy as far as the tips by applying the mutations (Algorithm 6). The genotype of the MRCA comes from the stationary distribution associated to the Markov chain of the mutational model. Figure 4 describes the process that generates the genotypes of the sample.

 ISSN: 2102-6238

A pure jump Markov process on the genealogy The evolution of the progeny of an individual over a lineage is a pure jump Markov process on $[\![a;b]\!]$. The Markov chain associated to the mutational model is the embedded chain of the jump process. Both have the same stationary distribution because the mutation rate (jump rate) does not depend on the allelic state. The ends of the paths of the process give the observed genotypes of the sample. (Figure 4). In Figure 4, each lineage corresponds to one path of the process. These paths are correlated and the dependence is described by the shared branches between the different lineages. Since the genotype of the MRCA comes from the stationary distribution, the marginal distribution of the genotype of any individual is also distributed according to this stationary distribution. However, the joint probability distribution of the genotypes of all individuals in the sample is more complex to describe. The correlation structure between the paths of the jump process is given by the branches shared by the lineages of the dendrogram. For example, in Figure 4, the Markov processes that explain the genotypes of individuals numbered 2 and 4 (represented in red and green respectively on the figure) share a large proportion of their trajectories, and are more correlated than, for instance, individuals numbered 2 and 5.

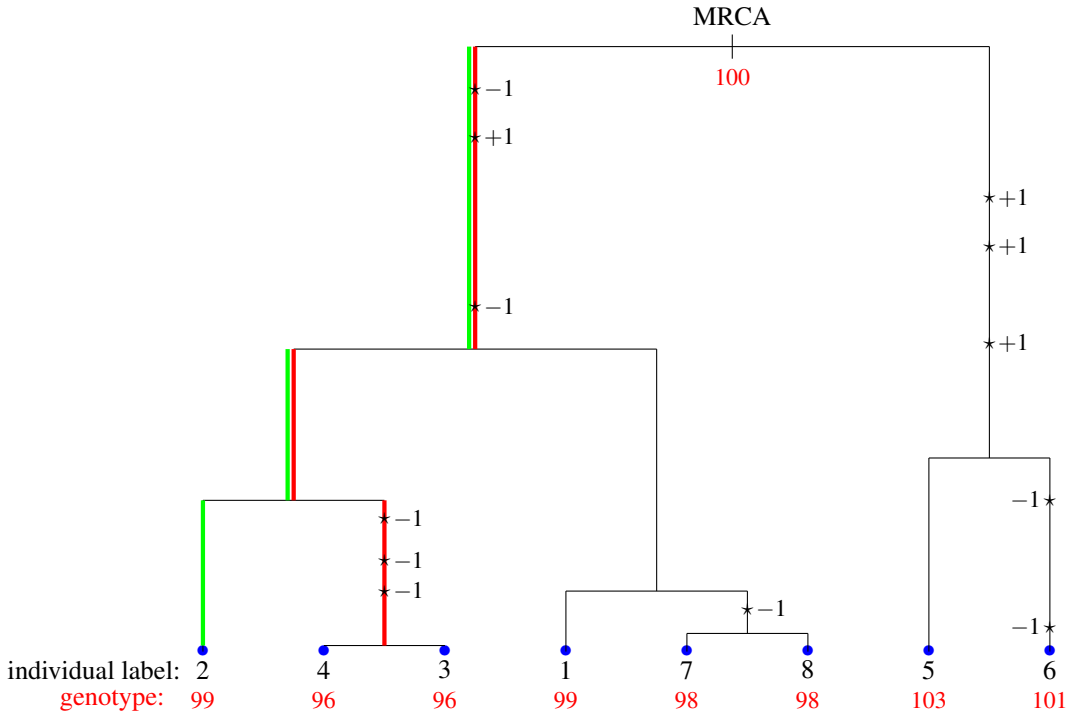

FIGURE 4. *Example of generating genotypes of a sample of* 8 *individuals in a microsatellite locus. The mutations are given by the SMM model. The mutation points are represented by stars (⋆) on lineages of the dendrogram. The genotype of an individual (represented by an integer in red on the dendrogram) is obtained by applying one-step transition of the Markov chain at each mutation point from the MRCA* (100) *genotype along his lineage.*

 ISSN: 2102-6238

3.2. *Mutations on sequence loci*

We now consider models on a sequence of DNA. A position within the locus is called a site.

Infinite site model This model is an approximation of a locus composed of a long DNA sequence that does not recombine, such that each site mutates very slowly. In this model, the number of sites is infinite and is often considered to be the whole interval $[0;1]$. At each site, only two values are possible, either 0 or 1, and the sequence of the MRCA is composed of 0's only. The positions of the mutations along the genealogy are drawn according to a Poisson point process of rate $\mu/2$ as above. At each position, a site is chosen uniformly at random on the interval $[0;1]$, and the value 0 at this site is mutated into 1.

Markov sequence models These models consider that the sequence of letters $\{A,C,G,T\}$ is of fixed length N. Each site evolves independently of the others. Along one branch of the genealogy, the site evolves according to a pure jump continuous time Markov process on the $\{A,C,G,T\}$-state space. It is often convenient to assume that the Markov process is time reversible. The simplest model that falls into this category is the model of Jukes et al. (1969), and the most complex one was set by Tamura and Nei (1993). In any case, the type of each site of the MRCA is drawn from the stationary distribution of the Markov process. On each branch of the genealogy, the site evolves as dictated by the Markov process, and at a given node, the process splits up into two independent processes.

4. Conclusion

Likelihood The likelihood of a given parameter $\phi = (\phi_{\text{dem}}, \phi_{\text{mut}})$ on the observed dataset $\mathbf{x}$ is therefore given by (1) and (2). The integral in (2) cannot be computed explicitly in a closed form, except in a very few simple cases. The space of $(\mathscr{G}_i, \mathscr{M}_i)$ satisfying $\mathscr{M}_i \to \mathbf{x}_i$ over which we have to integrate is a huge space of much greater dimension than the observed data $\mathbf{x}_i$ itself.

Ethier and Griffiths (1987); Griffiths (1989) developed recurrence formulas linking the likelihood $f(\mathbf{x}_i|\phi)$ of ϕ on the data to likelihoods $f(\mathbf{y}|\phi)$, of ϕ on samples $\mathbf{y}$ of smaller or equal size. De Iorio and Griffiths (2004a, equation (3)) describes the case of a closed population at equilibrium, and the case of several populations with migration De Iorio and Griffiths (2004b, equation (2)). Practically, it is not very efficient to evaluate the likelihood using this recurrence formula. Indeed, to obtain the value of $f(\mathbf{x}_i|\phi)$ for a given value of ϕ, the value of $f(\mathbf{y}|\phi)$ must be computed for all genetic data $\mathbf{y}$ on all samples smaller or equal to $\mathbf{x}_i$. Hence, the algorithm relying on the recurrence formula becomes exponentially more complex as the size of the sample increases, leading to what is known as the combinatorial explosion phenomenon.

Time scale To explain the observed dataset, we added a temporal dimension to set the latent process $(\mathscr{G}, \mathscr{M})$. Almost all coordinates of ϕ are set relatively to a given scale on this time axis. However, the dataset is collected at the present time and does not provide any information about the time scale. Indeed, applying a homothety of ratio λ on the time axis does not change the marginal distribution of the dataset $\mathbf{x}$ if we perform the following transformations on the components of ϕ:

— each event time parameter t is changed to λt,

 ISSN: 2102-6238

— each effective population size parameter Ne is changed to $\lambda\, Ne$,
— the mutation rate parameter μ is changed to μ/λ, and
— each migration rate parameter m is changed to m/λ.

When performing data analysis relying on classical statistics (frequentist), we face the problem of a lack of identifiability. This issue can be addressed by reparametrizing the model to eliminate the unknown scale on the time axis. To do so, one may introduce a reference effective population size, denoted Ne_{REF}, which is a linear combination of effective population sizes on a natural scale. The model then becomes identifiable if, first, we replace

— each data t_i by $\tau_i = t_i/Ne_{\text{REF}}$,
— each effective population size Ne_i by $\overline{Ne_i} = Ne_i/Ne_{\text{REF}}$,
— each mutation rate μ_i by $\theta_i = 4Ne_{\text{REF}}\,\mu_i$ and
— each migration rate m_{ij} by $\overline{m_{ij}} = Ne_{\text{REF}}\, m_{ij}$,
— etc.

and, remove one effective population size in ϕ. Commonly, Ne_{REF} corresponds to the sum of effective population sizes of the observed populations. For example, in Figure 1, we could set $Ne_{\text{REF}} = Ne_1 + Ne_2 + Ne_3 + Ne_4$, use the transformation described above and remove $\overline{Ne_1}$. For biologists, the interpretation of the identifiable parameterization requires using a *rule of thumb* to set the time scale. However, Bayesian analysis allows this problem to be circumvented since a posterior distribution is well defined, even if we lack identifiability. Moreover, information on the time scale can be set in the prior. This make it possible to introduce some variability on the value of this scale through a directly interpretable distribution on the parameters. This is an argument in favor of Bayesian analysis in population genetics, which was clearly highlighted by Beaumont and Rannala (2004).

Likelihood approximation There are three main families of algorithms.

— The first family is based on a data augmentation MCMC that samples from the distribution of $(\phi, \mathscr{G}, \mathscr{M})$ knowing $\mathbf{x}$. The space we have to sample in this case is of very high dimension. Hence, the running time can be very long, and the chain can easily get stuck in a part of the high dimensional space we have to sample.
— The second family is based on importance sampling. The idea here is to introduce an auxiliary distribution commonly called *proposal* distribution to sample the space of interest and introduce a weight that corrects the discrepancy between the target distribution and the proposal. In this context, the calculation of (1) and (2) is an integral according to a probability distribution. Calibration of the proposal is required to control the Monte Carlo error, bu this is difficult. Ad hoc strategies can be used to calibrate the proposal in this context Stephens and Donnelly (2000); De Iorio and Griffiths (2004a,b); De Iorio et al. (2005).
— The third family involves Approximate Bayesian Computation (ABC) methods that use many simulated datasets according to the model instead of likelihood calculation. Thus, ABC methods Beaumont et al. (2002); Marjoram et al. (2003); Marin et al. (2011) compare several simulated datasets to the observed dataset using summary statistics which are supposed to be informative for the posterior calculation. The target of these Monte Carlo methods is a degraded version of the posterior: this is the distribution of the parameters where the summary statistics are known instead of the complete dataset. This family pro-

 ISSN: 2102-6238

vides less relevant results than importance sampling to approximate the likelihood function at each point of interest in the parameter space. However, these methods are much more flexible because they need only to simulate according to the model and provide important information through summary statistics.

Software There are several available tools in the literature but they do not cover all the questions encountered in practice. Here are some representative examples of simulation and estimation software.

— *Ms* of Hudson (2002) simulates demographic scenarios with migration between populations using Kingman's coalescent.

— *IBDSim* of Leblois et al. (2009) simulates scenarios with migration (isolation by distance) in discrete time, generation by generation, and samples the latest generation.

— *DIYABC* of Cornuet et al. (2008) simulates demographic scenarios involving divergence and admixture events without migration. Note that *DIYABC* provides procedures of estimation and model selection based on ABC.

— *Migraine* of Rousset and Leblois (2012) computes likelihood surfaces using an importance sampling algorithm.

— *IM, IMa* of Nielsen and Wakeley (2001); Hey and Nielsen (2004, 2007) analyzes demographic models with migration. Both Bayesian and frequentist approaches are included. Posterior distribution and likelihood surfaces may be approximated with MCMC methods.

References

Beaumont, M., Zhang, W., and Balding, D. (2002). Approximate Bayesian Computation in population genetics. *Genetics*, 162:2025–2035.

Beaumont, M. A. and Rannala, B. (2004). The Bayesian revolution in genetics. *Nature Reviews Genetics*, 5(4):251–261.

Cornuet, J. M., Beaumont, M. A., Estoup, A., and Solignac, M. (2006). Inference on microsatellite mutation processes in the invasive mite, Varroa destructor, using reversible jump Markov chain Monte Carlo. *Theoretical Population Biology*, 69(2):129–144.

Cornuet, J.-M., Santos, F., Beaumont, M. A., Robert, C. P., Marin, J.-M., Balding, D. J., Guillemaud, T., and Estoup, A. (2008). Inferring population history with DIYABC: a user-friendly approach to Approximate Bayesian Computation. *Bioinformatics*, 24(23):2713–2719.

De Iorio, M. and Griffiths, R. C. (2004a). Importance sampling on coalescent histories. I. *Advances in Applied Probability*, 36(2):417–433.

De Iorio, M. and Griffiths, R. C. (2004b). Importance sampling on coalescent histories. II:Subdivided population models. *Advances in Applied Probability*, 36(2):434–454.

De Iorio, M., Griffiths, R. C., Leblois, R., and Rousset, F. (2005). Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theoretical Population Biology*, 68(1):41–53.

Ethier, S. N. and Griffiths, R. C. (1987). The Infinitely-Many-Sites Model as a Measure-Valued Diffusion. *The Annals of Probability*, 15(2):515–545.

Griffiths, R. C. (1989). Genealogical-tree probabilities in the infinitely-many-site model. *Journal of mathematical biology*, 27(6):667–680.

Griffiths, R. C. and Marjoram, P. (1997). An ancestral recombination graph. *Institute for Mathematics and its Applications*, 87:257.

Hey, J. and Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. *Genetics*, 167(2):747–760.

Hey, J. and Nielsen, R. (2007). Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences*, 104(8):2785–2790.

 ISSN: 2102-6238

Hudson, R. R. (2001). Two-locus sampling distributions and their application. *Genetics*, 159(4):1805–1817.
Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338.
Jukes, T. H., Cantor, C. R., et al. (1969). Evolution of protein molecules. *Mammalian protein metabolism*, 3(21):132.
Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626.
Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press.
Kingman, J. F. C. (1982a). The coalescent. *Stochastic Processes and their Applications*, 13:235–248.
Kingman, J. F. C. (1982b). Exchangeability and the Evolution of Large Populations. In Koch, G. and Spizzichino, F., editors, *Exchangeability in Probability and Statistics*, pages 97–112. North-Holland, Amsterdam.
Kingman, J. F. C. (1982c). On the Genealogy of Large Populations. *Journal of Applied Probability*, 19:27–43.
Leblois, R., Estoup, A., and Rousset, F. (2009). IBDSim: A computer program to simulate genotype data under Isolation By Distance. *Molecular Ecology Ressources*, 9(1):107–109.
Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2011). Approximate Bayesian computational methods. *Statistics and Computing*.
Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.
Müller, N. F., Rasmussen, D. A., and Stadler, T. (2017). The structured coalescent and its approximations. *Molecular biology and evolution*, 34(11):2970–2981.
Nielsen, R. and Wakeley, J. (2001). Distinguishing Migration From Isolation: A Markov Chain Monte Carlo Approach. *Genetics*, 158(2):885–896.
Rousset, F. and Leblois, R. (2012). Likelihood-Based Inferences under Isolation by Distance: Two-Dimensional Habitats and Confidence Intervals. *Molecular Biology and Evolution*, 29(3):957–973.
Sainudiin, R., Stadler, T., and Véber, A. (2015). Finding the best resolution for the kingman–tajima coalescent: theory and applications. *Journal of mathematical biology*, 70(6):1207–1247.
Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):605–635.
Stumpf, M. P. and McVean, G. A. (2003). Estimating recombination rates from population-genetic data. *Nature reviews. Genetics*, 4(12):959–968.
Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees. *Molecular biology and evolution*, 10(3):512–526.
Wakeley, J. (2005). *Coalescent Theory: An Introduction*. Roberts & Company Publishers.
Whittaker, J. C., Harbord, R. M., Boxall, N., Mackay, I., Dawson, G., and Sibly, R. M. (2003). Likelihood-based estimation of microsatellite mutation rates. *Genetics*, 164(2):781–787.

 ISSN: 2102-6238