Peer Community Journal

Section: Mathematical & Computational Biology

Population genetics: coalescence rate and demographic parameters inference

Olivier Mazet [ID],[1] and Camille Noûs[2]

Volume **3** (2023), article e53

https://doi.org/10.24072/pcjournal.285

Abstract

Inferring the demographic history of species is a great challenge in population genetics. This history is classically represented as a history of size changes, ignoring population structure. We present here the work carried out over the last decade around the concept of IICR (Inverse Instantaneous Coalescence Rate), which makes it possible to link, on the one hand, the history of the true population size for a panmictic population, and the inferred size, sometimes called "effective size", when structure is taken into account. We show that population structure can lead to misinterpretations of some demographic history inference results, we propose a framework for inferring structure-specific demographic parameters (number and size of subpopulations, migration rates), and we analyze the link between IICR and some form of selection modeling on genetic sequences.

[1]Université de Toulouse, Institut National des Sciences Appliquées de Toulouse, Institut de Mathématiques de Toulouse, Toulouse, France, [2]Laboratoire Cogitamus, France https://www.cogitamus.fr/

Contents

1. Introduction

Note to the reader: this article was originally conceived for readers who are mathematicians, with the objective of presenting an example of the application of mathematics in life sciences. It is for them that the first part has been designed, which readers already in the know can advantageously skip. On the other hand, the few mathematical formulas are likely to make the reading more difficult for people with a biology background, we hope that the explanations of their meanings will remain accessible to most people. The presentation of results in a field that is essentially multidisciplinary is always a little perilous.

We propose in this article a brief description of the work, over almost a decade, resulting from a collaboration between mathematicians and biologists from four different research laboratories, identifiable as the co-authors of the articles whose results are described here, and implicitely co-authors of this article, under the signature of Camille Noûs. This modeling work is part of population genetics, and is therefore essentially at the interface between mathematical tools, more particularly probabilistic ones, and biological data, more specifically genetic ones.

In a first part, we briefly present the theory of coalescence, which is the basis of our models, and the problems that this modeling tries to address. In a second part we describe our first results and the development of the IICR (Inverse of Instantaneous Coalescence Rate), the nodal point from which the different research paths we have been following are branching. The first results of these paths are summarized in the two following parts, one about the inference of demographic parameters of a structured population, the other about the consideration of selection.

2. Coalescent theory in population genetics

Population genetics is the study of the evolution of the genotypes in a population of living beings, under various evolutionary pressures such as mutation, selection or genetic drift. Initiated in the first half of the XXth century with the work of the British statistician Ronald Fisher and the American geneticist Sewall Wright, it has seen the emergence of a backward model called **coalescent**, the first developments of which are due to the British mathematician John Kingman in the 80's (Kingman, 1982).

The coalescent theory consists in sampling individuals – more precisely loci of individuals' genomes – in the present population, and tracing their genealogies back in time, until a common ancestor, from two or more lineages, is found. The instants, backward to the past, when such common ancestors appear are called **coalescence times**, and are considered as random variables with values in $\mathbb{N}$ or in $\mathbb{R}^+$.

The mathematical object of interest is then the joint distribution of the various coalescence times of family trees, which allows to express the observable quantities in the genomes of a present population, as functions of this distribution. Those functions depend on genetic parameters (like mutation rate, recombination rate, selection rate) and demographic parameters (like sizes and numbers of sub-populations, migration rates between sub-populations). Note that in

this paper, we will focus on loci separated by recombination events, and the topology of each tree of each locus won't matter. Observations of genetic sequences can then be used to infer genetic or demographic parameters using various statistical methods, and technological advances over the last two decades have made it possible to acquire huge masses of data, which can be used to refine existing models and develop new ones.

Wright-Fisher model and Kingman coalescent.

More precisely, the Wright-Fisher model describes the evolution of a population of $2N$ individuals (the individuals can be genes or loci) with the following assumptions: in each generation, each individual independently generates a number of descendants following a Poisson distribution of the same constant parameter, with the offsprings completely replacing their parents, all conditioned by the fact that the size of the population must remain constant. The process can be described in an equivalent way backward in time: each individual of a given generation randomly chooses its parent in the previous generation in a uniform way. An illustration of the process is given in Figure 1 (Figure from Hein et al. (2004)).
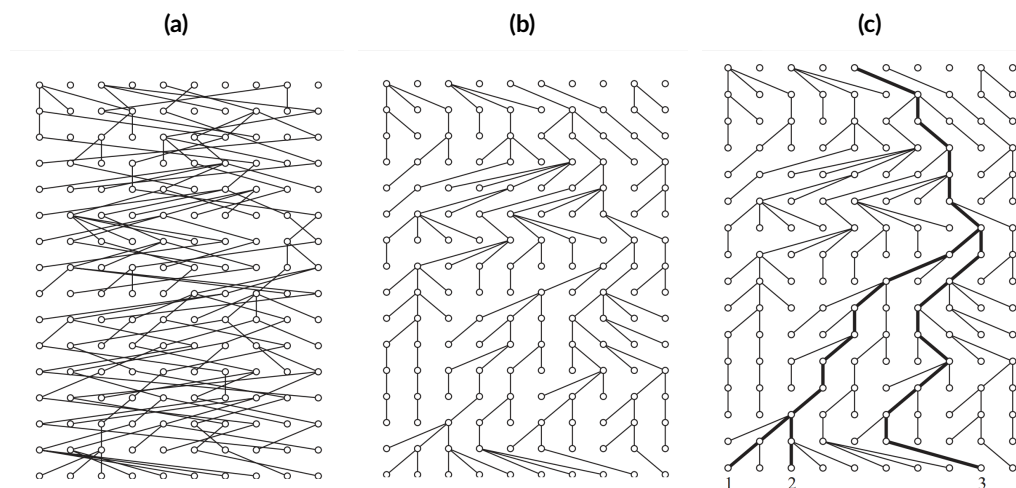
Figure 1 – Here is a realization of the Wright-Fisher process on 16 generations with a population size of $2N = 10$. Panel **(a)** presents the evolution when each row is a generation, the individuals have been rearranged in panel **(b)** in order to highlight the family tree, and for panel **(c)** three individuals have been chosen in the last generation, their respective lineages having been put in bold. We see that the first coalescence between individuals 1 and 2 takes place two generations ago in the past, and that the last coalescence to the most recent common ancestor of individuals 1, 2 and 3 takes place nine generations ago in the past.

If we now consider a pair of individuals in the last generation, and if we note T_2 the waiting time for the coalescence of the two lineages (going back in time), we have

$$\mathbb{P}(T_2 > i) = \left(1 - \frac{1}{2N}\right)^i,$$

and if we suppose N to be large, by changing the time scale, we obtain the usual approximation of the geometric distribution by the exponential distribution

$$\mathbb{P}(T_2 > 2Nt) = \left(1 - \frac{1}{2N}\right)^{2Nt} \sim e^{-t}.$$

We can thus generalize for the first renormalized coalescence time, which for convenience we keep noting T_k, even if it is not expressed in the same units, for k individuals sampled in the last generation:

$$\mathbb{P}(T_k > t) \sim e^{-\binom{k}{2}t},$$

and one thus obtains the coalescence tree, corollary of Kingman's coalescent, where the successive times of coalescence are independent, following exponential distributions of parameters equal to the binomial coefficients $\binom{k}{2}$.

Demographic complications of the coalescence model.

The coalescent defined this way is only valid in the context of a so-called panmictic population, i.e. without geographical structure (each individual randomly chooses its parent in the whole population), and with constant size. We will see how to generalize the coalescent in the case of a population of changing size, and in the case of a structured population.

Population size change. If we consider that the size of the population can vary, by posing $N(i)$ the size of the population at generation i in the past, and by considering the quantity

$$\lambda(t) = \frac{N(\lfloor 2Nt \rfloor)}{N(0)}, \tag{1}$$

the relative size of the population in the past with the same temporal renormalization as in the previous section, it can be shown (see e.g. Tavaré (2004), section 2.4) that under reasonable conditions of variation of λ in the neighbourhood of infinity, the coalescence time T_k of k individuals sampled in the present satisfies

$$\mathbb{P}(T_k > t) = \exp\left(-\binom{k}{2}\int_0^t \frac{\mathrm{d}\tau}{\lambda(\tau)}\right). \tag{2}$$

But unlike the panmictic case, the successive T_k are no longer independent, which makes the global study of the tree more difficult.

Structured population. To relax the assumption of panmixy opens the door to multiple ways of modeling population structuring. Classically, the global population is considered to be made up of subpopulations (called islands, or demes), each of which is panmictic, between which migration events may occur, with rates that may depend on each pair of islands. The demographic parameters of the model are thus: the number of islands n, the respective renormalized sizes $(s_i)_{i=1\ldots n}$ of the n islands (again assumed constant in time), and the renormalized migration rates (to take into account the scaling already described which allows us to go to continuous time by assuming that populations are sufficiently large) $(M_{ij})_{i\neq j}$ between the islands i and j.

The description of the coalescent tree thus becomes much more complex, but the information can be summarized, as Hilde Herbots-Wilkinson showed in 1994 in her landmark thesis work (Herbots, 1994). If we note $\alpha = (\alpha_1, \ldots, \alpha_n)$ the configuration where α_i represents the number of lineages present in the island i, then the coalescence process can be described by the infinitesimal generator Q such that

$$Q(n_\alpha, n_\beta) = \begin{cases} \alpha_i \frac{M_{ij}}{2} & \text{if } \beta = \alpha - \epsilon^i + \epsilon^j \quad (i \neq j) \\ \frac{1}{s_i}\frac{\alpha_i(\alpha_i-1)}{2} & \text{if } \beta = \alpha - \epsilon^i \\ -\sum_i \left(\alpha_i\frac{M_i}{2} + \frac{1}{s_i}\frac{\alpha_i(\alpha_i-1)}{2}\right) & \text{if } \beta = \alpha \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where ϵ^i is the vector of size n whose components are all zero except the i-th which is 1. In order to have positive integers as indices of the matrix, we denote by n_α and n_β the respective numbers of the α and β configurations, once chosen a prior ordering of all possible configurations. The first line stands for a migration event, from island i to island j, with the corresponding migration rate $\alpha_i\frac{M_{ij}}{2}$. The second means that two lineages in island i coalesced, with rate $\frac{1}{s_i}\frac{\alpha_i(\alpha_i-1)}{2}$.

**Genetic parameters, estimations and inferences.**

From a genetic point of view, all these models are assumed to be neutral, i.e. not taking into account the possible influence of selection in the reproductive capacity of each individual. However, it is possible to easily incorporate the phenomena of mutation and recombination into these models, because they can be considered as events independent of the genealogical process. The

classical assumptions are that each mutation or recombination event affects a different part of the genome (the so-called *infinite site model*), and that the mutation and recombination rates are constant both in time and along the genetic sequences.

Mutation and genetic diversity. Mutation events are distributed on the genealogical tree according to a Poisson process, and we can link genetic diversity data, by observing for example the number of alleles of a given gene, or its distribution, and more generally the quantification of polymorphism, with the configuration of the tree (topology, length of branches) associated with the chosen model. By choosing a model, we can estimate the mutation parameter, and on the contrary, by assuming the mutation parameter to be known, we can estimate the lengths of the branches of the tree, and thus have information on the distributions of the coalescence times. Among the best known estimators, let us mention Watterson's θ_W based on the number of segregating sites (Watterson, 1975) or the number of pairwise nuleotide differences (Tajima, 1983).

Recombination and Sequential Markovian Coalescence. The phenomenon of recombination is much more difficult to incorporate into these models than mutation, since it requires sexual reproduction, and at each recombination event the resulting genome is derived not from one parent but from two, thus exponentially increasing the number of ancestors involved for each lineage of individuals sampled in the present population. The ancestral recombination graph (ARG, see Griffiths and Marjoram (1996)) requires a computational treatment that is very quickly prohibitive when the sample size increases.

The work of McVean and Cardin (2005) allowed, under an original hypothesis of the Markovian dependence property *along the genome* (hence the so-called *sequential* property), to greatly restrict the space to be explored for statistical inference methods. Several demographic parameter inference software packages then emerged, including the famous PSMC (for *Pairwise Sequentially Markovian Coalescent*, Li and Durbin (2011)), which has been widely used since 2011, and allows to estimate the variation of the population size (noted $\lambda(t)$ in equation (1)), using the genetic data from a diploid individual only (fully sequenced genome), see for example Figure 2.
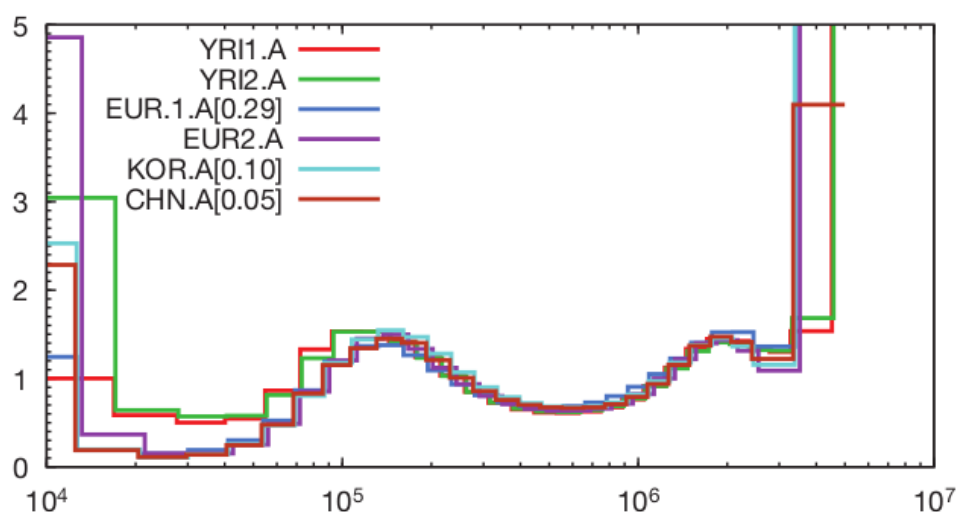
Figure 2 – Demographic inference obtained from human DNA from individuals of different populations (Li and Durbin, 2011). On the x-axis, the number of years in the past. On the y-axis, the renormalized size of the population, assumed to be panmictic.

3. Consideration of population structure, central role of the IICR

Some work in the first decade of this century (Chikhi et al., 2001, 2010; Wakeley, 2001) highlighted the effect that a structured population could generate on statistics assuming panmixia, with notably the detection of false bottleneck signals in some cases. This can be problematic, for

instance for conservation population issues. It is not always easy to quantify some kind of degree of structuration, there exists a large bibliography on this topic (see for instance Chakraborty (1993) or Excoffier (2004)).

After a preliminary study which consisted in analyzing a simple case of comparison of an island model and a size change model (Mazet et al., 2015), we obtained a first result, which has become a nodal point of our subsequent research and which we present here.

The IICR and the change in size. We have highlighted in Mazet et al. (2016) the following result. Considering that whatever the chosen demographic model is, the coalescence time T_2 of the lineages of two individuals chosen in the present population is a random variable with values in $\mathbb{R}^+$. This variable thus can be considered as a lifetime of density f_{T_2}, and as such, admits a "failure rate" which here translates into an **(instantaneous) rate of coalescence** equal to

$$\mu(t) = \frac{f_{T_2}(t)}{\mathbb{P}(T_2 > t)}.$$

The density of T_2 can thus always be written

$$f_{T_2}(t) = \mu(t)\exp\left(-\int_0^t \mu(\tau)\mathrm{d}\tau\right),$$

hence

(4)
$$\mathbb{P}(T_2 > t) = \exp\left(-\int_0^t \mu(\tau)\mathrm{d}\tau\right).$$

If we now bring equation (4) together with the particular case $k = 2$ of equation (2), we realize that in the panmictic case, the change in size $\lambda(t)$ is exactly equal to $\frac{1}{\mu(t)}$, which is thus the inverse of the instantaneous coalescence rate, noted by the acronym **IICR**. Two important consequences can be drawn from this observation:

(1) The sole data of the T_2 distribution cannot be informative of the demographic model, if we don't know if it is structured or not, since whatever it is, there is always a panmictic model which will provide exactly this T_2 distribution. Indeed, it is sufficient to choose the inverse of the coalescence rate as the size change.

(2) What software like PSMC infers, on data from a single diploid genome and however long it may be, is the IICR associated with the demographic model, which is usually **not the change in size** of the population when it is structured.

Taking the second consequence further, as a proof of concept we built a constant size demographic model based solely on a symmetric island model, with the number of islands also constant, where only the migration parameter is allowed to vary. As we can see in Figure 3 extracted from Mazet et al. (2016), the PSMC output on data simulated under this model is very similar to that on real human data.

There is obviously no question of claiming that the human population is structured in symmetrical islands and that its population size has remained constant over the course of evolution, but this example prompts us to question the interpretation of the IICR, which is the object inferred by the PSMC, and shows that it is necessary to investigate further, before drawing any conclusions about the demographic history of a population.

Links with previous results. Beyond the formal framework, which confirms what simulations in Wakeley (2001), Chikhi et al. (2001) or Chikhi et al. (2010) had already shown, i.e. bottleneck or population expansion signals depending on whether the lineages are sampled in the same demes or not (see also Peter et al. (2010) or Heller et al. (2013)), the IICR also makes it possible to establish links with works on various concepts of effective sizes of a structured population.

Indeed, the IICR of a stationary structured model, after a transitory phase which depends on the sampling, always converges towards a plateau which could be considered as an "asymptotic" effective size and whose value is calculable in an exact way, according to the demographic parameters (number and size of the demes, rate of migration between the demes). It turns out that for the n-island and for fairly large migration values, in the first approximation this value is equal
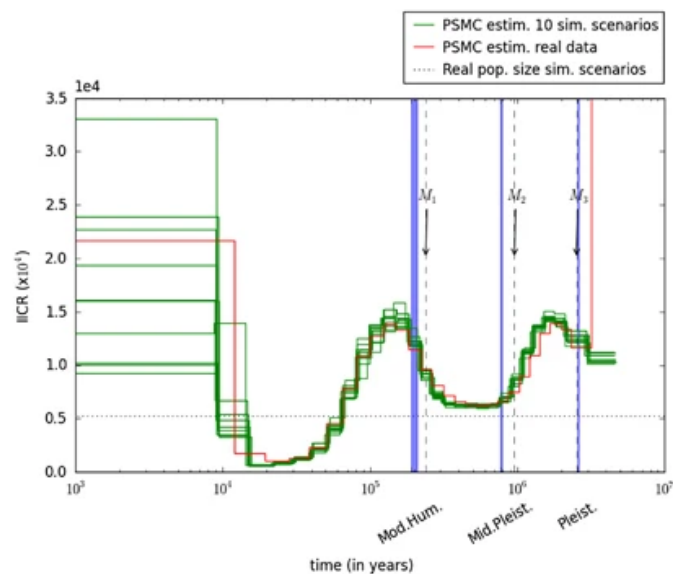
Figure 3 – In red the PSMC of real data from a human (CHN.A in Figure 2). In green the PSMC of 10 simulations of the same model in islands, of constant size, with three changes of migration rate represented by the vertical dotted lines, at 2.52 Myr ago, 0.95 Myr ago and 0.24 Myr ago. The blue shaded areas correspond to the beginning of the Pleistocene at 2.57–2.60 Myr ago, the beginning of the Middle Pleistocene at 0.77–0.79 Myr ago and the oldest known fossils of anatomically modern humans at 195–198 kyr ago. Following Li and Durbin (2011), we assumed that the mutation rate was $\mu = 2.5 \times 10^{-8}$ and that generation time was 25 years. We also kept their ratio between mutation and recombination rates. Each deme had a size of 530 diploids and the total number of haploid genomes was thus constant and equal to 10 600.

to the metapopulation size Nn (n being the number of subpopulations and N the size of each of them), and in the second approximation (by adding the next term of the Taylor expansion) this value is equal to $N(n + \frac{(n-1)^2}{nM})$, which is the effective size computed in Nei and Takahata (1993) for this model.

Let us also note in the work of Whitlock and Barton (1997) a certain similarity of approach, but for the case where the unit of time remains the generation, and where the *eigenvalue effective population size* defined by Ewens (1982) is studied in connection with "their" Inverse of Instantaneous Coalescence Rate.

Influence of the sample size. From the first consequence drawn above, we explored theoretically what the data of a third lineage could bring as additional information. We then showed that in the simple case of a population structured in islands, then adding the information of the T_3 distribution to the T_2 distribution is enough to distinguish this model from the panmictic model having the same T_2 distribution, thus the same IICR (Grusea et al., 2018).

This result provides theoretical evidence that a sample size strictly greater than two is sufficient to distinguish an island structured model from a panmictic model, but initial attempts to move into practice have not yet been successful, because of the precision required, which often blends into the noise of the real data.

Structure and IICR: sampling strategy. Exploratory work was then done (Chikhi et al., 2018), using simulated data, to find out what signatures are left on the IICR produced by different types of structured models, and thus indirectly (or directly when dealing with software of the same type as PSMC) on the false signals of size changes that these models generate. As an illustration, we present in Figure 4 extracted from this paper, the simulated IICRs in a model with three islands and asymmetric migration rates, by sampling a diploid individual in each of the islands. We see that not only the structure of the model can give false signals of size change for software assuming panmixia, but also the IICR is **dependent on the sampling location**, for the same

demographic history. Indeed, sampling a diploid individual in island 3 would provide a signal of monotonic population decrease, whereas sampling it in island 1 or 2 would provide a signal of first expansion, followed by reduction, with different magnitudes. This finding also deserves to be further explored for use in model selection.
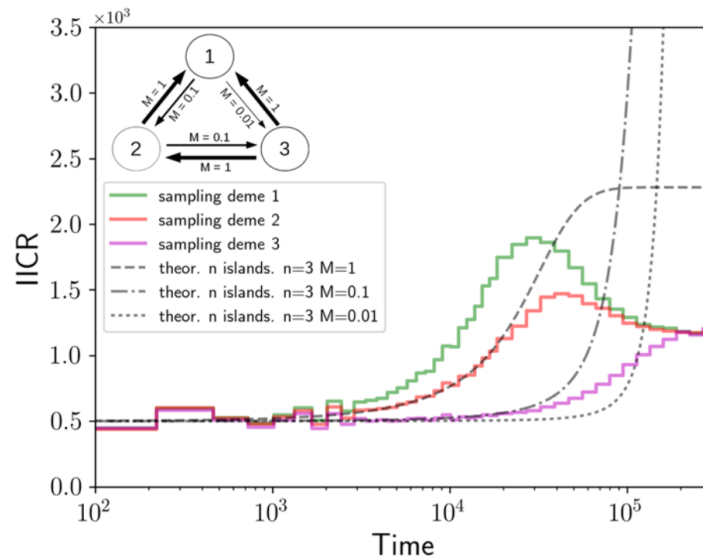
Figure 4 – The size of each island is constant. However, the IICRs of each pair of sequences are time-varying functions, and these functions even may not be monotonic. Furthermore, they differ depending on the island from which the pair is sampled. This Figure comes from Chikhi et al. (2018).

The IICR as a model validation. The IICR can also be used as a **summary statistic** of a given model, for validation or rejection. In the same paper (Chikhi et al., 2018) we thus tested a number of models proposed in the literature for human evolution, for *Homo sapiens* as well as for *Homo neandertalensis*. Simulating the IICR of some of these models allowed us to discard them, as the IICR produced differed radically from that estimated by PSMC on human data. For example the Figure 5 illustrates this situation for the models proposed in Yang et al. (2012), which was supposed to prove that admixture between humans and Neanderthals explained better the data than ancient structure.

IICR and structured coalescent. On the theoretical side, work has been undertaken to calculate, as precisely as we like, the IICR of any structured model (Rodríguez et al., 2018). The modeling initiated by Herbots provides a set of infinitesimal generators of Markov processes (see formula (3)), and it is possible to exploit the semi-group property of the exponentials of these matrices. Indeed, changes in some parameters of the structured models, such as island sizes or migration rates, leave the state space of the process unchanged, so the matrices can be piecewise constant functions of time. For example, if we suppose that at a date T in the past some of the parameters M_{ij} or s_i change, and if we note by Q_0 the generator for the time $0 \leq t \leq T$ and by Q_1 the one corresponding to the time $t > T$, the transition semi-group of the Markov chain can be written as follows:

$$P_t = \begin{cases} e^{tQ_0}, & \text{if } t \leq T \\ e^{TQ_0}e^{(t-T)Q_1}, & \text{otherwise}. \end{cases}$$

In particular, the distribution of T_2^α, coalescence time of two lineages starting from a α configuration, is deduced from

$$\mathbb{P}(T_k^\alpha \leq t) = P_t(n_\alpha, n_c),$$
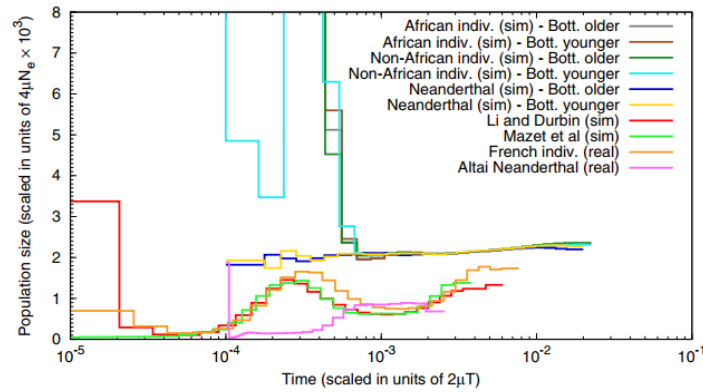
Figure 5 – The PSMCs of real modern human and Neanderthal data (the last two in the order of the legend), set against the PSMCs of simulated data with parameter values used by the authors of different proposed models in Yang et al. (2012) : the African, Non-African and Neanderthal individuals were simulated under the model of recent admixture with a bottleneck that was either older or younger than the admixture event. And for the record, the panmictic model suggested by Li and Durbin (2011) and the fictional structured model proposed by Mazet et al. (2016). For more detailed explanation, see Chikhi et al. (2018).

where n_α is the number of the state corresponding to α, and n_c the number of the coalescence state. Its density is then equal to $f_{T_k^\alpha}(t) = P'_t(n_\alpha, n_c)$, where

$$P'_t = \begin{cases} e^{tQ_0}Q_0, & \text{if } t < T \\ e^{TQ_0}e^{(t-T)Q_1}Q_1, & \text{otherwise}. \end{cases}$$
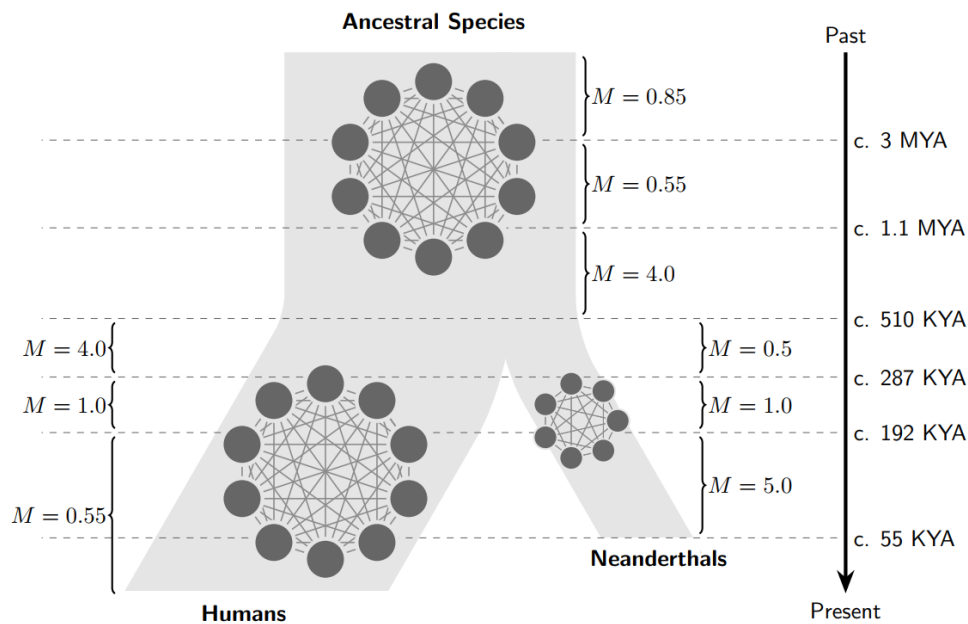
These explanations allow to numerically determine the theoretical IICRs of a large number of structured models, such as the continent-islands model in Figure 6, with possible changes in demographic parameters, such as subpopulation sizes or migration rates.
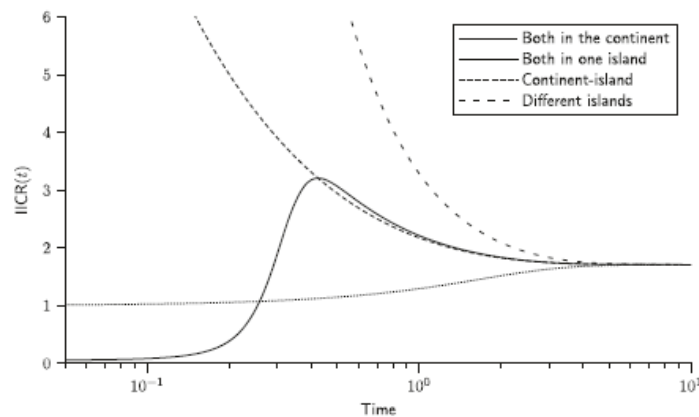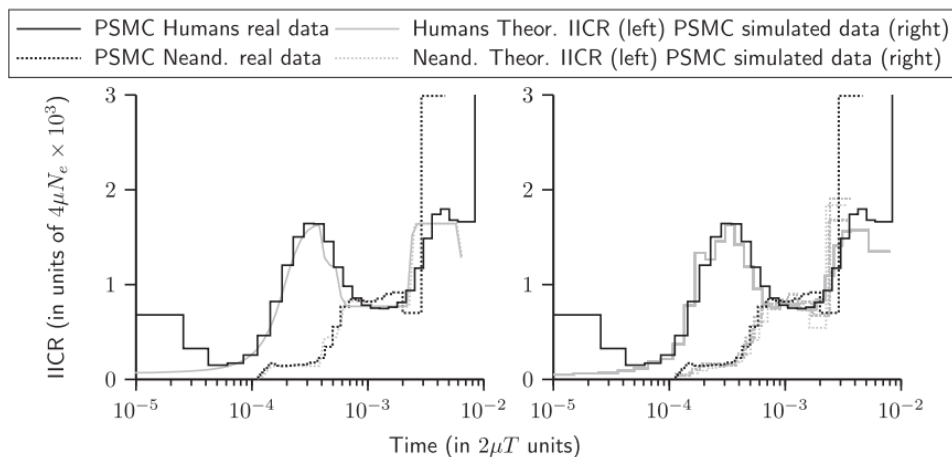

Figure 6 – Theoretical IICR of the structured model with a continent of size 1, three islands of sizes $\frac{1}{20}$, and migration rates proportional to sizes between islands and the continent (no migration between islands). We find, as in the simulations in Chikhi et al. (2018), the importance of sampling location, as well as the obvious false signals of population size changes that software such as PSMC could infer, here the demographic model being constant over time.

Also as a proof of concept, an extended fictional model of human evolution was proposed, integrating Neanderthals alongside modern humans in the same constant size structured model, with only the migration coefficients allowed to change. The model is described in Figure 7, and the simulated PSMCs are presented in Figure 8 with PSMCs of real data. We then see that relatively simple structured model can explains the PSMCs without change of the metapopulation

sizes before and after the split between Neanderthals and modern humans from an unknows *Homo* species.

Figure 7 – Hypothetical scenario presenting humans and Neanderthals as structured species derived from an unknown Homo species that was itself structured, and without any gene flow between the two species after the split. The times at which gene flow (M) changed are indicated by horizontal lines.

Figure 8 – Superposition of the PSMCs of real Neanderthal and Sapiens data, with the theoretical IICRs of the proposed structured model (left) and the PSMCs of the simulated data from this structured model (right).

4. Inference of parameters in a structured model

IICR of T_2 for a structured model.

The theoretical possibility (presented in Rodríguez et al. (2018)) of numerically computing the IICR of two sampled lineages in any structured model, as a function of the demographic parameters such as the number of islands, the successive island sizes, the successive migration rates, and the times of change of parameters like sizes or migration rates, opens the way to estimate these parameters from IICRs inferred from real data, e.g. via the inevitable PSMC. The challenges of

complexity and computation time, even in the simplest case of the symmetric island model, have been overcome thanks to the work of Armando Arredondo, part of his PhD thesis (Arredondo Soto, 2021), with the design and realization of a software for inferring such parameters, called SNIF (Structured Non-stationary Inferential Framework), presented in Arredondo et al. (2021).

Testing this software on simulated data revealed a first problem of identifiability between subpopulation sizes and migration rates. Second, if we want to obtain an acceptable level of precision for the estimation of the migration rates, the number of different values over time (this number is called the number of "components") should not be too large, generally not more than 5 or 6. On the other hand, the estimated number of subpopulations is extremely reliable. A synthesis of these first results can be seen on Figure 9.
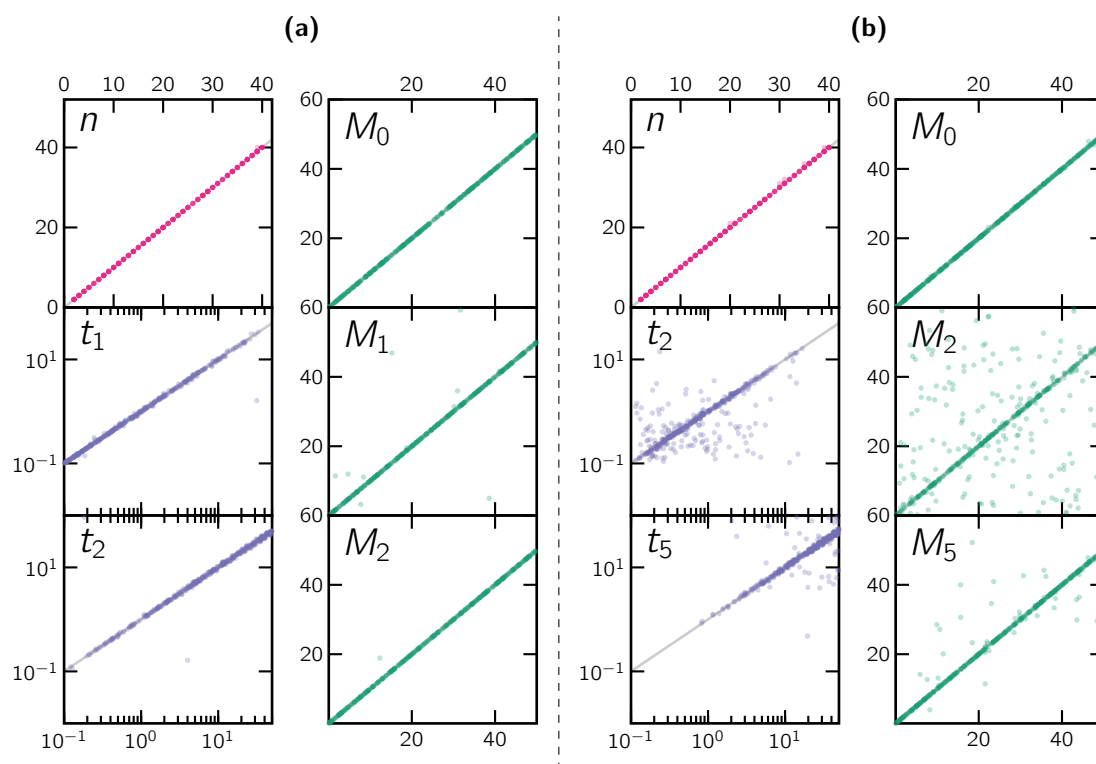

Figure 9 – Scatter plots of simulated and inferred parameters. n is the number of islands, t_i the i-th change of value of the migration parameter, and M_i the i-th value of the latter. Panel **(a)** corresponds to scenarios with $c = 3$ components, and **(b)** to scenarios with $c = 6$ components. The different sub-panels represent the simulated (horizontal axis) versus inferred (vertical axis) parameter values for all the parameters (or a representative selection of parameters in the case of panel **(b)**) of $L = 400$ unscaled simulated scenarios.

An application on human data also allows to find a good model in symmetrical islands which explains surprisingly well the graph produced by the PSMC, see Figure 10.

This method has already been used to contribute to the study of the evolution of species of microcebes, Malagasy lemuriform primates (*Microcebus murinus* and *Microcebus ravelobensis*). The results are published in (Teixeira et al., 2021). Other data on other species of mammals are being analysed using this software.

Increase of the sample size.

A natural way to increase the precision of the estimation of demographic parameters, is natural to increase the size of the statistical sample. For instance, the distribution of alleles frequency over the gene, generally called SFS (for *Site Frequency Spectrum*), is commonly used in population genetics, since it's easy to extract values from real data. Nevertheless, the average SFS is theoretically known only for a panmictic model (see for example Griffiths and Tavaré (1998)), but the calculation becomes of great combinatorial complexity for any structured model. In the case of
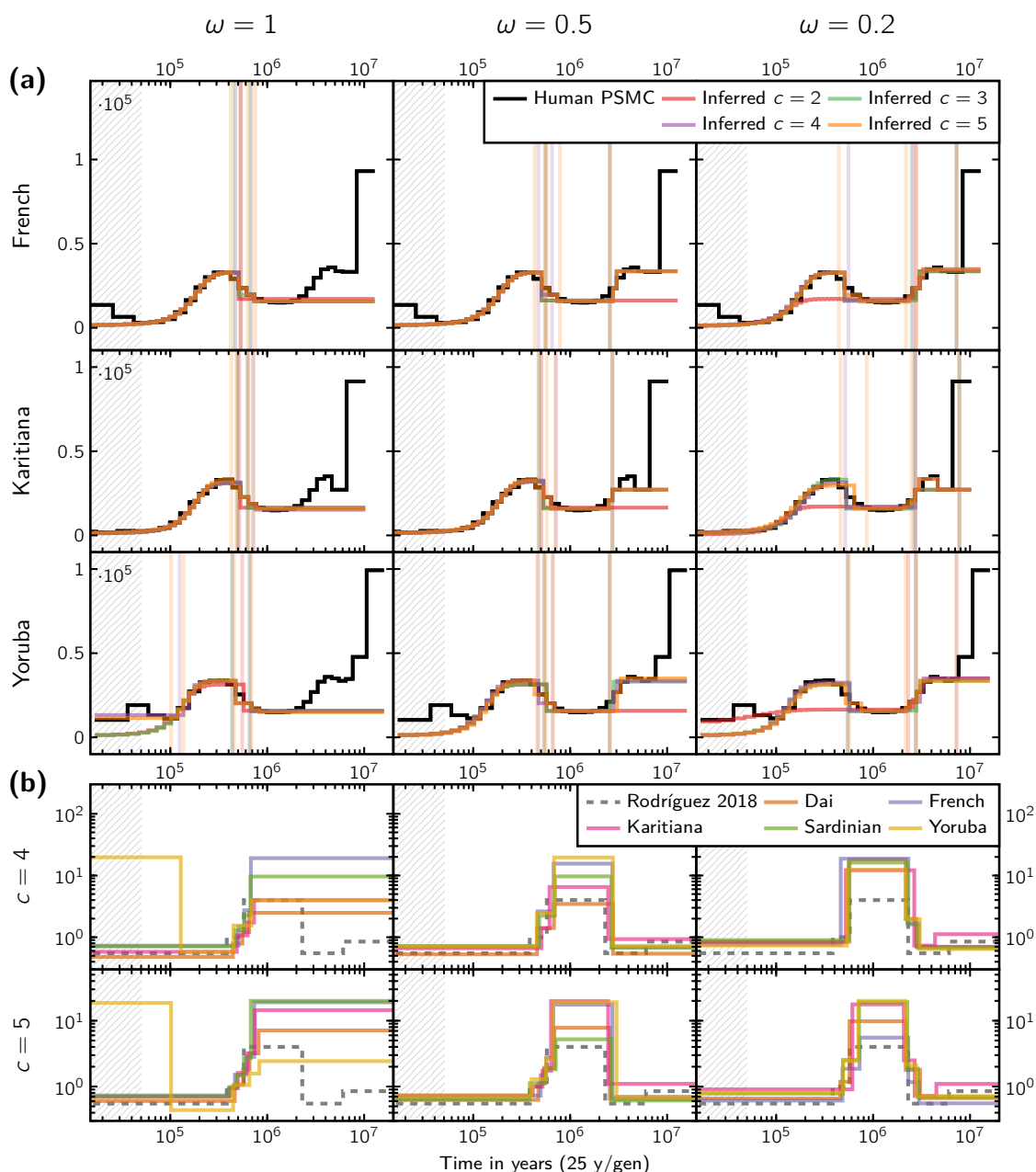
Figure 10 – Results of performing demographic inference on three representative human PSMC curves. Panel **(a)** shows the various IICR plots inferred for the different populations, numbers of components c and weight parameters ω used, together with the target IICR curves (or PSMC plots) on which these estimations are based. Panel **(b)** shows the connectivity graphs for the same set of inferred scenario. As a reference point, the connectivity graph of the scenario proposed in Rodríguez et al. (2018) is also shown. The vertical axes represent migration rates (M).

the island model, Armando Arredondo has just completed the theoretical treatment of obtaining the average SFS for any value of k, as well as the feasibility in computation time for a sample size of $k \leq 26$ in the current state of computing capabilities (Arredondo Soto, 2021, chapter 3). It now remains to implement this algorithm in the inference software.

IICR$_k$: derivation and first results. Coming back to the coalescences times T_k of such a sample of size k, their IICR (noted here IICR$_k$) for $k > 2$ is theoretically easily computable thanks to the infinitesimal generator of equation (3) and the extensions exposed in section 3. Indeed, the IICR$_k$ of the first time T_k of coalescence of k lineages can be defined in the same way as the

IICR (which is in fact the IICR_2):

$$\text{IICR}_k(t) = \frac{f_{T_k}(t)}{\mathbb{P}(T_k > t)}.$$

While we know that in the panmictic case we have

$$\forall k \geq 2, \forall t > 0, \qquad \text{IICR}_k(t) = \frac{1}{\binom{k}{2}}\text{IICR}_2(t),$$

this is not the case for a structured model (as we formally showed for the symmetric island model in Grusea et al. (2018)). There already exist powerful methods to estimate the IICR_k of real genomic data of sample size k, notably the extensions of the PSMC, called MSMC (for *Multiple Sequentially Markov Coalescent*, see Schiffels and Durbin (2013)). The practical problem comes from the fact that the larger the sample size, the shorter the coalescence time, and thus the fewer the genomic traces on the data, because the number of mutation and recombination events decreases very quickly, and falls below the acceptable threshold for the statistical estimation to be satisfactory.

Still, for small values of k, (let's say from $k = 3$ to $k = 8$), some results are being obtained in a work in progress, which show that IICR_k could provide an efficient tool to distinguish a structured population from a panmictic one.

5. IICR and consideration of selection

All the models we have discussed so far are so-called neutral models, i.e. they do not take into account the selection pressure that individuals undergo at some loci to increase the frequency of new advantageous alleles (positive selection), decrease the frequency of new deleterious alleles (negative selection) or maintain polymorphism (balancing selection). A large bibliography deals with the detection of positive selection events (the so-called selective sweeps see the review in Walsh and Lynch (2018)), and many studies have investigated the influence of demography on sweep detection (including for instance Jensen et al. (2005) or Bonhomme et al. (2010), but there are many others). On an other side, fewer and more recent papers studied the impact of selection on demography inference, like Ewing and Jensen (2016) or Johri et al. (2020) for negative selection, and for sweeps, Schrider et al. (2016) or Harris and Rogers (2020).

One classical way to model the long-term combined effect of selection and genetic linkage (i.e. linked selection) on genomic sequences is to assume that some portions of the genome undergo a recurrent higher impact of selection due to their local gene content or recombination rate, resulting in an effective size that differs from that in neutral areas (see Hill and Robertson (1966), Charlesworth (2009), Gossmann et al. (2011) or Jiménez-Mena, Bataillon, et al. (2016) for example). Under this approach, regions under positive or negative selection are modelled by a lower effective size, while regions under balancing selection are modelled by a higher. This is a way to make the coalescence rate variable over the genome, this rate being linked to the reproductive capacity.

Since the IICR is directly related to the coalescence rate, it is natural to explore the influence of modeling selection by the variability of the effective size along the genome on the IICR (Boitard et al., 2022). A theoretical calculation allows us to show, under the simple hypothesis of a panmictic population, that if we assume the existence of K classes of the genome under respective effective sizes equal to $\lambda_i = \frac{1}{\mu_i}$ for $i = 1 \ldots K$, each corresponding to a proportion a_i of the genome (with $\sum_i^K a_i = 1$) then the IICR is

$$\text{IICR}(t) = \frac{\sum\limits_{i=1}^{K} a_i e^{-\mu_i t}}{\sum\limits_{i=1}^{K} a_i \mu_i e^{-\mu_i t}},$$

and a basic calculation indicates that for all values of K, a_i and λ_i, this IICR is always increasing on $\mathbb{R}^+$, with

$$\text{IICR}(0) = \frac{1}{\sum\limits_{i=1}^{K} a_i \mu_i} \quad \text{and} \quad \lim_{t \to +\infty} \text{IICR}(t) = \max_{i=1\ldots K} \lambda_i.$$

It is thus mathematically shown that the IICR will exhibit a decline signal of population size, as it can be seen in Figure 11. An interpretation could be that in recent past the coalescent rate is high (and consequently the effective size appears to be small) because of coalescent events happening in zones of small effective size, and in distant past are remaining the coalescent events occuring in zones of large effective size. Figure 11 shows also that under these assumptions, the largest effective size present in the genome, even under a very small proportion, has a significant influence on the growth of the IICR as a function of time from the present to the past. The magnitude of this variation may be surprising, but the simulations also show, especially in the presence of structure in the population, that it may be statistically difficult to detect the final plateau, given the small number of corresponding coalescence events.
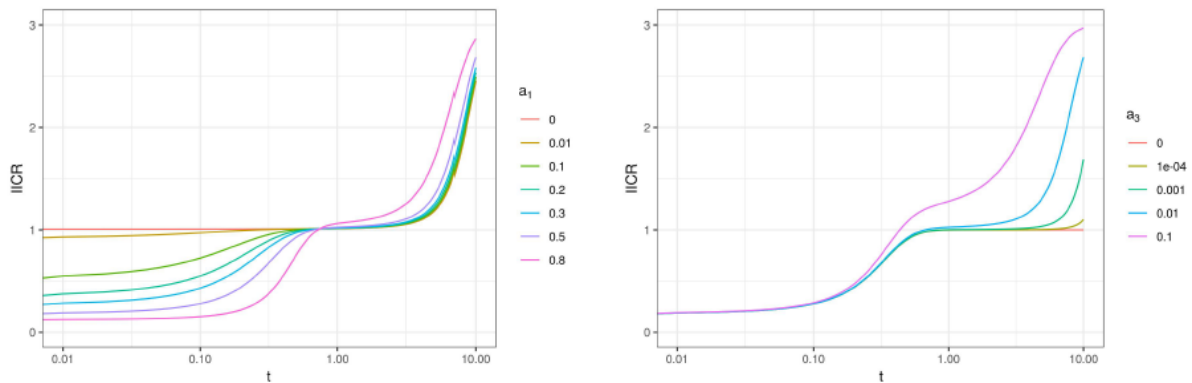
Figure 11 – Example of IICR for $K = 3$, $\lambda_1 = 0.1$, $\lambda_2 = 1$ and $\lambda_3 = 3$. On the left we set $a_3 = 0.01$ and on the right $a_1 = 0.5$. The value λ_3 determines the limit, and a_3 the speed of convergence, with more or less pronounced transient plateaus depending on the other values.

It is possible to generalize the derivation of the analytic expression of the IICR if the population is structured. If we denote by $f_i(t)$ and by a_i respectively the density of the coalescence time T_2^i and the proportion of the i-th of the K classes of the genome, let's recall that in the panmictic case we have $\mathbb{P}(T_2^i > t) = e^{-\mu_i t}$ and $f_i(t) = \mu_i e^{-\mu_i t}$, so if the population is not panmictic, the formula becomes

$$\text{IICR}(t) = \frac{\sum_{i=1}^{K} a_i \mathbb{P}(T_2^i > t)}{\sum_{i=1}^{K} a_i f_i(t)},$$

and thanks to our previous work on island-structured models, we can combine the effects of structure and of selection and numerically calculate the corresponding IICRs (see Figure 12). We can then see that, even if we find the same monotonic pattern as in the panmictic case, the structure hides partly the growth towards the limit value, which is reached quite far in the past.

Throughout this section, the coalescence rate was considered variable along the genome, but was assumed to be constant over time to model the long-term average effect of selection. The IICR mathematical framework also allows to model *transient selection*, which was done in Boitard et al. (2022) for one specific scenario, again from the point of view of the impact on the PSMC. In order not to overload this section, we let the interested reader refer to this article.

Finally, it should be noted that this study only uses information from the T_2 distribution, which may explain why our results contradict some of the literature by showing signals of recent population decline. Indeed, for larger sample sizes, especially when using the SFS, negative selection leads to an excess of singletons, which under the neutral hypothesis leads to a signature of a recent population expansion.
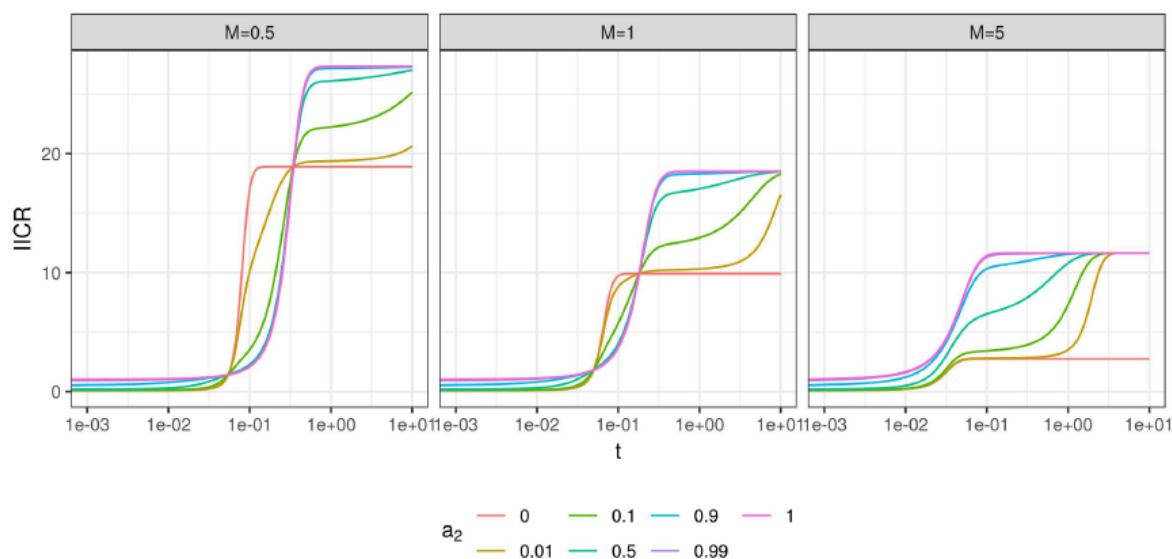
Figure 12 – Example of IICR for a symmetric 10-island model, migration rate M, and a genome with $K = 2$ effective size classes $\lambda_1 = 0.1$ and $\lambda_2 = 1$ of relative proportions a_1 and a_2.

6. Conclusion and prospects

In summary, the IICR of T_2, despite its intrinsic limitations, on the one hand because it is a distribution of a variable which is not directly observable, and on the other hand because it is based on a sample of size 2, proves to be an extremely fertile object of modeling. The matrix writing of this time function (it should be noted that the mathematical objects studied in our work are part of a general formalism coming from the theory of phase-type distributions, see for example Hobolth et al. (2019)) facilitates precise numerical calculations (Rodríguez et al., 2018), and makes accessible powerful inference methods, such as the recently developed SNIF program (Arredondo et al., 2021).

The IICR also sheds new light on a concept with which it is naturally associated, that of *effective size*, which is the source of an abundant literature (see, for example Charlesworth (2009), or the recent Waples (2022) for a complete review), and is subject to many different, sometimes contradictory, interpretations. The starting point is the evidence of a direct correspondence between the IICR and the population size in panmictic condition. But a first level of hypothesis complexification, with the introduction of a population structuring, quickly leads to erroneous conclusions about size changes. We can thus observe variations in the effective size that do not correspond to those of the real size, or even that are in the opposite sense (Chikhi et al., 2018; Rodríguez et al., 2018). It is nevertheless difficult to quantify a level of structuration from which the problem highlighted could be qualified as serious. Moreover, it seems obvious that if structure is omnipresent, so is real variation in population size, and the combination of the two phenomena, as well as their variability over time, makes the inference of demographic parameters and the interpretation of results more complex. Similarly, it is useful to know how to detect possible effective size variations, induced by the introduction of genomic areas under selection, positive or negative (Boitard et al., 2022). Another crucial contribution of the IICR is to highlight the importance of the sampling strategy (Chikhi et al., 2018), the exploitation of which should allow new methods for model selection.

Among the other more immediate prospects, we note the continuation of the theoretical study of the IICR for models a little more elaborate than the symmetric island model (first of all the asymmetric island model, the island-continents model, or even the one or two dimensional stepping-stone model), with the objective of highlighting the influence of the values of the demographic parameters on the variations of the IICR, via the eigenvalues of the associated infinitesimal generator. Finally, in order to increase the predictive and explanatory capacities of our models by enlarging the sample size, we have to develop the inference methods from the

existing one (Arredondo et al., 2021), by incorporating on the one hand the average SFS of an island model, and on the other hand the IICR of T_k.

Acknowledgements

Fundings

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

7. References

Arredondo A, Mourato B, Nguyen K, Boitard S, Rodríguez W, Noûs C, Mazet O, Chikhi L (2021). *Inferring number of populations and changes in connectivity under the n-island model. Heredity* **126**, 896–912. https://doi.org/10.1038/s41437-021-00426-9.

Arredondo Soto A (2021). *Inférence d'histoires démographiques de populations structurées et application à l'évolution humaine.* PhD thesis. Toulouse, INSA.

Boitard S, Arredondo A, Chikhi L, Mazet O (2022). *Heterogeneity in effective size across the genome: effects on the Inverse Instantaneous Coalescence Rate (IICR) and implications for demographic inference under linked selection. Genetics, iyac008.* https://doi.org/10.1093/genetics/iyac008.

Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, Blott S, SanCristobal M (2010). *Detecting selection in population trees: the Lewontin and Krakauer test extended. Genetics* **186**, 241–262. https://doi.org/10.1534/genetics.110.117275.

Chakraborty R (1993). *Analysis of genetic structure of populations: meaning, methods, and implications. Human population genetics: A centennial tribute to JBS Haldane*, 189–206. https://doi.org/10.1007/978-1-4615-2970-5_14.

Charlesworth B (2009). *Effective population size and patterns of molecular evolution and variation. Nature Reviews Genetics* **10**, 195–205. https://doi.org/10.1038/nrg2526.

Chikhi L, Rodríguez W, Grusea S, Santos P, Boitard S, Mazet O (2018). *The IICR (inverse instantaneous coalescence rate) as a summary of genomic diversity: insights into demographic inference and model choice. Heredity* **120**, 13–24. https://doi.org/10.1038/s41437-017-0005-6.

Chikhi L, Bruford MW, Beaumont MA (2001). *Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. Genetics* **158**, 1347–1362. https://doi.org/10.1093/genetics/158.3.1347.

Chikhi L, Sousa VC, Luisi P, Goossens B, Beaumont MA (2010). *The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. Genetics* **186**, 983–995. https://doi.org/10.1534/genetics.110.118661.

Ewens W (1982). *On the concept of the effective population size. Theoretical Population Biology* **21**, 373–378. `https://doi.org/10.1016/0040-5809(82)90024-7`.

Ewing GB, Jensen JD (2016). *The consequences of not accounting for background selection in demographic inference. Molecular ecology* **25**, 135–141. `https://doi.org/10.1111/mec.13390`.

Excoffier L (2004). *Analysis of population subdivision. Handbook of statistical genetics*. `https://doi.org/10.1002/0470022620.bbc25`.

Gossmann TI, Woolfit M, Eyre-Walker A (2011). *Quantifying the variation in the effective population size within a genome. Genetics* **189**, 1389–1402. `https://doi.org/10.1534/genetics.111.132654`.

Griffiths R, Marjoram P (1996). *An ancestral recombination graph, pp. 257–270 in IMA Volume on Mathematical Population Genetics, edited by Donnelly P., Tavaré S.*

Griffiths RC, Tavaré S (1998). *The age of a mutation in a general coalescent tree. Stochastic Models* **14**, 273–295. `https://doi.org/10.1080/15326349808807471`.

Grusea S, Rodríguez W, Pinchon D, Chikhi L, Boitard S, Mazet O (2018). *Coalescence times for three genes provide sufficient information to distinguish population structure from population size changes. Journal of Mathematical Biology* **78**, 189–224. `https://doi.org/10.1007/s00285-018-1272-4`.

Harris NS, Rogers AR (2020). *Genomic regions linked to soft sweeps approximate neutrality when inferring population history from site pattern frequencies. bioRxiv*, 2020–04. `https://doi.org/10.1101/2020.04.21.053629`.

Hein J, Schierup M, Wiuf C (2004). *Gene genealogies, variation and evolution: a primer in coalescent theory.* Oxford University Press, USA.

Heller R, Chikhi L, Siegismund HR (2013). *The confounding effect of population structure on Bayesian skyline plot inferences of demographic history. PloS one* **8**, e62992. `https://doi.org/10.1371/journal.pone.0062992`.

Herbots HMJD (1994). *Stochastic models in population genetics: genealogy and genetic differentiation in structured populations.* PhD thesis.

Hill WG, Robertson A (1966). *The effect of linkage on limits to artificial selection. Genetics Research* **8**, 269–294. `https://doi.org/10.1017/S0016672300010156`.

Hobolth A, Siri-Jegousse A, Bladt M (2019). *Phase-type distributions in population genetics. Theoretical Population Biology* **127**, 16–32. `https://doi.org/10.1016/j.tpb.2019.02.001`.

Jensen JD, Kim Y, DuMont VB, Aquadro CF, Bustamante CD (2005). *Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics* **170**, 1401–1410. `https://doi.org/10.1534/genetics.104.038224`.

Jiménez-Mena B, Bataillon T, et al. (2016). *Heterogeneity in effective population size and its implications in conservation genetics and animal breeding. Conservation genetics resources* **8**, 35–41. `https://doi.org/10.1007/s12686-015-0508-5`.

Johri P, Riall K, Jensen JD (2020). *The impact of purifying and background selection on the inference of population history: problems and prospects. BioRxiv.* `https://doi.org/10.1093/molbev/msab050`.

Kingman JFC (1982). *The coalescent. Stochastic processes and their applications* **13**, 235–248. `https://doi.org/10.1016/0304-4149(82)90011-4`.

Li H, Durbin R (2011). *Inference of human population history from individual whole-genome sequences. Nature* **475**, 493–496. `https://doi.org/10.1038/nature10231`.

Mazet O, Rodríguez W, Chikhi L (2015). *Demographic inference using genetic data from a single individual: Separating population size variation from population structure. Theoretical Population Biology* **104**, 46–58. `https://doi.org/10.1016/j.tpb.2015.06.003`.

Mazet O, Rodríguez W, Grusea S, Boitard S, Chikhi L (2016). *On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? Heredity* **116**, 362. `https://doi.org/10.1038/hdy.2015.104`.

McVean GA, Cardin NJ (2005). *Approximating the coalescent with recombination. Philosophical Transactions of the Royal Society B: Biological Sciences* **360**, 1387–1393. `https://doi.org/10.1098/rstb.2005.1673`.

Nei M, Takahata N (1993). *Effective population size, genetic diversity, and coalescence time in subdivided populations. Journal of Molecular Evolution* **37**, 240–244. https://doi.org/10.1007/BF00175500.

Peter BM, Wegmann D, Excoffier L (2010). *Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. Molecular ecology* **19**, 4648–4660. https://doi.org/10.1111/j.1365-294X.2010.04783.x.

Rodríguez W, Mazet O, Grusea S, Arredondo A, Corujo JM, Boitard S, Chikhi L (2018). *The IICR and the non-stationary structured coalescent: towards demographic inference with arbitrary changes in population structure. Heredity* **121**, 663–678. https://doi.org/10.1038/s41437-018-0148-0.

Rogers AR (2023). *Estimates of Effective Population Size in Subdivided Populations. Peer Community in Mathematical and Computational Biology*, 100150. https://doi.org/10.24072/pci.mcb.100150.

Schiffels S, Durbin R (2013). *Inferring human population size and separation history from multiple genome sequences. Nature Genetics* **8**, 919–925. https://doi.org/10.1038/ng.3015.

Schrider DR, Shanku AG, Kern AD (2016). *Effects of linked selective sweeps on demographic inference and model selection. Genetics* **204**, 1207–1223. https://doi.org/10.1534/genetics.116.190223.

Tajima F (1983). *Evolutionary relationship of DNA sequences in finite populations. Genetics* **105**, 437–460. https://doi.org/10.1093/genetics/105.2.437.

Tavaré S (2004). *Part I: Ancestral inference in population genetics*. In: *Lectures on probability theory and statistics*. Springer, pp. 1–188. https://doi.org/10.1007/978-3-540-39874-5_1.

Teixeira H, Salmona J, Arredondo A, Mourato B, Manzi S, Rakotondravony R, Mazet O, Chikhi L, Metzger J, Radespiel U (2021). *Impact of model assumptions on demographic inferences – the case study of two sympatric mouse lemurs in northwestern Madagascar. BMC Ecol Evo* **21**. https://doi.org/10.1186/s12862-021-01929-z.

Wakeley J (2001). *The coalescent in an island model of population subdivision with variation among demes. Theoretical population biology* **59**, 133–144. https://doi.org/10.1006/tpbi.2000.1495.

Walsh B, Lynch M (2018). *Evolution and selection of quantitative traits*. Oxford University Press. https://doi.org/10.1093/oso/9780198830870.001.0001.

Waples RS (2022). *What is N_e, anyway? Journal of Heredity* **113**, 371–379. https://doi.org/10.1093/jhered/esac023.

Watterson G (1975). *On the number of segregating sites in genetical models without recombination. Theoretical population biology* **7**, 256–276. https://doi.org/10.1016/0040-5809(75)90020-9.

Whitlock MC, Barton NH (1997). *The effective size of a subdivided population. Genetics* **146**, 427–441. https://doi.org/10.1093/genetics/146.1.427.

Yang MA, Malaspinas AS, Durand EY, Slatkin M (2012). *Ancient structure in Africa unlikely to explain Neanderthal and non-African genetic similarity. Molecular biology and evolution* **29**, 2987–2995. https://doi.org/10.1093/molbev/mss117.