Heterogeneity in effective size across the genome: effects on the Inverse Instantaneous Coalescence Rate (IICR) and implications for demographic inference under linked selection

Simon Boitard*, Armando Arredondo[†], Camille Noûs[‡], Lounès Chikhi[§,**], Olivier Mazet[†]

*: CBGP, Université de Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France.

[†]: Université de Toulouse, Institut National des Sciences Appliquées, Institut de Mathématiques de Toulouse, Toulouse, France.

[‡]: Laboratoire Cogitamus, Toulouse, France.

[§]: Instituto Gulbenkian de Ciência, Oeiras, Portugal

**: Laboratoire Évolution & Diversité Biologique (EDB UMR 5174), CNRS, IRD, UPS, Université de Toulouse Midi-Pyrénées, Toulouse, France

Running title:

The IICR under linked selection

Keywords:

demographic inference, linked selection, effective population size, coalescence times, population structure, drosophila melanogaster, humans

Corresponding author:

Simon Boitard

CBGP, 755 avenue du Campus Agropolis, CS 30016, 34988 Montferrier sur Lez cedex, France

simon.boitard@inrae.fr

0033 4 99 62 33 36

Abstract

The relative contribution of selection and neutrality in shaping species genetic diversity is one of the most central and controversial questions in evolutionary theory. Genomic data provide growing evidence that linked selection, i.e. the modification of genetic diversity at neutral sites through linkage with selected sites, might be pervasive over the genome. Several studies proposed that linked selection could be modelled as first approximation by a local reduction (e.g. purifying selection, selective sweeps) or increase (e.g. balancing selection) of effective population size (N_e). At the genome-wide scale, this leads to a large variance of N_e from one region to another, reflecting the heterogeneity of selective constraints and recombination rates between regions. We investigate here the consequences of this variation of N_e on the genome-wide distribution of coalescence times. The underlying motivation concerns the impact of linked selection on demographic inference, because the distribution of coalescence times is at the heart of several important demographic inference approaches. Using the concept of Inverse Instantaneous Coalescence Rate, we demonstrate that in a panmictic population, linked selection always results in a spurious apparent decrease of N_e along time. Balancing selection has a particularly large effect, even when it concerns a very small part of the genome. We quantify the expected magnitude of the spurious decrease of N_e in humans and *Drosophila melanogaster*, based on N_e distributions inferred from real data in these species. We also find that the effect of linked selection can be significantly reduced by that of population structure.

Introduction

One of the greatest challenges of evolutionary biology is to understand how natural selection, mutation, recombination and genetic drift have shaped and are still shaping the patterns of genomic diversity of species living today (Charlesworth, 2010, Lewontin, 1974, Walsh and Lynch, 2018). In the last decade genomic data have become increasingly available for both model and non-model species. It is expected that by analysing these genomic data we will be able to better understand the respective roles of mutation and recombination and characterize the relative importance of drift and selection in evolutionary processes (Charlesworth, 2010, Lewontin, 1974). In particular, it is believed that we will be able to identify the regions that have been shaped by selection, and those that may be more neutral (Johri et al., 2020, Pouyet et al., 2018). The relative importance of selection and neutrality in generating the genomic patterns of diversity we see today has been at the heart of many evolutionary debates and controversies over the last decades (Kimura, 1983, Lewontin, 1974, Ohta, 1992) and recent studies suggest that it still is (Comeron, 2017, Jensen et al., 2019, Kern and Hahn, 2018).

The concept of effective size (N_e) is central to these debates (Charlesworth, 2009) because selection is expected to be more efficient when N_e is large, and genetic drift to be the main driver of evolutionary change when N_e is small (Ohta, 1992). For instance, Charlesworth (2009) notes that it is expected that an autosomal locus under positive selection will behave neutrally when $s < 1/4N_e$, where s is the selection intensity at this locus. At the same time it is commonly assumed that selection will itself imply a variation of N_e across the genome (Charlesworth, 2009, Gossmann et al., 2011, Jiménez-Mena et al., 2016b). For instance, Gossmann et al. (2011) write that "*The effective population size is expected to vary across the genome as a consequence of genetic hitchhiking (Smith and Haigh, 1974) and background selection (Charlesworth et al., 1993)*". They add that "*The action of both positive and negative natural selection, is expected to reduce the effective population size leading to lower levels of genetic diversity and reduced effectiveness of selection.*" They also stress that "*The evidence that there is variation in N_e within a* *genome comes from three sources. First, it has been shown that levels of neutral genetic diversity are correlated to rates of recombination in Drosophila [...], humans [...], and some plant species...*". In his 2009 review on the concept of N_e Charlesworth (2009) made a similar comment: "*N_e may also vary across different locations in the genome of a species [...] because of the effects of selection at one site in the genome on the behaviour of variants at nearby sites*". More recently, Jiménez-Mena et al. (2016a) stated that "*recent studies [...] suggest that different segments of the genome might undergo different rates of genetic drift, potentially* ***challenging the idea that a single*** N_e ***can account for the evolution of the genome***" (emphasis ours).

Under these explicit or implicit modelling frameworks, genomic regions with limited genetic diversity are thus seen as regions of low N_e as a result of selective sweeps (Smith and Haigh, 1974) or background selection (Charlesworth et al., 1993), whereas regions with very high levels of genetic diversity may be seen as regions of large N_e and could be explained by balancing selection (Charlesworth, 2009) (see also Hill and Robertson (1966)). Following that rationale, Jiménez-Mena et al. (2016b) suggested that different species might thus differ in the statistical distribution of N_e across the genome and they presented such distributions for eleven species.

Given the central role played by the N_e concept to detect, identify, and even *conceptualize* selection, it may be important, perhaps even enlightening, to explore the consequences of the ideas presented above with the concept of IICR (inverse instantaneous coalescence rate) recently introduced by Mazet et al. (2016). Indeed, the IICR is equivalent to the past temporal trajectory of N_e, previously defined as the coalescent N_e (Sjödin et al., 2005), in a panmictic population under neutrality, and it is the quantity estimated by the popular PSMC method of Li and Durbin (2011). The IICR was first defined by Mazet et al. (2016) for a sample size of two and its properties were studied under several models of population structure (Chikhi et al., 2018, Grusea et al., 2018, Rodríguez et al., 2018) and it has been shown that it can be used for demographic inference under neutrality and models of population structure Arredondo et al. (2021), Chikhi et al. (2018). These studies showed that the IICR will significantly change over time when populations are structured, even when population size is actually constant. They also outlined that the IICR not only depends on the model of population structure but also on the sampling scheme, which questions the notion that an N_e can be easily associated to (or is a property of) the model of interest when the model is structured (Chikhi et al., 2018, Rodríguez et al., 2018). The reason for this dependency is that the IICR is by definition a function of the distribution of coalescence times for two genes (T_2), which is itself a function of both the evolutionary model and the location (in time and space) of the sampled genes.

One important assumption of the IICR studies mentioned above is that this distribution of T_2 is homogeneous along the genome. The IICR, as defined and computed in previous studies, is thus a genomic average assuming that all loci follow a single Wright-Fisher model with or without population structure but with the same number of haploid genes. Whichever definition of N_e one assumes, the underlying model assumes that N_e is constant along the genome. If we now assume that there is an N_e that varies across the genome as a consequence of selection (even as an approximation) then the IICR should be a function of the underlying distribution of these N_e values across the sampled genes. Genomic regions under different selection regimes might then exhibit specific signatures leading to differing IICR curves for each region. Alternatively, these regions might not be easy to identify but they might still influence the average genomic IICR estimated from sequenced genomes. In the present study we thus wish to explore ideas related to drift, selection and patterns of genomic diversity by studying the consequences of this putative genomic variation of N_e on the IICR.

We first study the IICR under panmixia and constant population size but assuming that N_e can vary across the genome, using hypothetical distributions of N_e and distributions inferred from genomic data. We then generalise the model to integrate population structure or population size variation through time. In particular we consider a structured model in which we allow several classes of genomic regions to evolve under different n-island models each characterized by a specific deme size. We also apply this approach to a structured model inferred for humans, which includes temporal variations of the migration rate between demes. Finally, we consider temporal variations of the genomic N_e to model possible transitory effects of selection under panmixia. Altogether, we advocate the use of the IICR as a concept that may help clarify what N_e means and as one way, among others, to improve our understanding of the recent and ancient evolutionary history of species.

The IICR under panmixia with several classes of (constant size) N_e along the genome

Model

We assume that the genome can be divided in K independent classes, each of them characterized by a different N_e that is constant over time. To model these differences of N_e, we consider that each class i ($i = 1 \dots K$) evolves under a constant size Wright-Fisher (WF) model (*i.e.* panmictic with non-overlapping generations) with diploid population size $\lambda_i N$ (2 $\lambda_i N$ haploids), for some reference population size N corresponding to the actual number of diploids. Note that $2N$ represents an actual number of haploid genomes and that under the WF model, there is no ambiguity and N represents the N_e under neutrality. Thus, λ_i reflects the ratio of effective population size N_e in class i relative to N and for convenience we may sometimes refer to λ_i as *the* effective population size in class i. Assuming that N is large (i.e. that all $\lambda_i N$ are large), we rescale time by units of $2N$ generations and study the coalescence process resulting from this model. If we sample in the present (at time $t = 0$) k genes for a locus from the i^{th} class of the genome, the genealogy of this sample will follow a standard coalescent model with coalescence rate $\mu_i = \frac{1}{\lambda_i}$ between any pair of sequences. This model is associated to the successive coalescence times T_j^i, $j = k \dots 2$, where T_j^i is the coalescence time of j lineages sampled in the class i.

We will study the properties of the IICR under this model. As a reminder, the IICR was originally defined for a sample of size $k = 2$ by Mazet et al. (2016), who showed how it could be computed for any model for which T_2 values could be generated (*i.e.* any model under the coalescent possibly involving complex population structure and population size changes).

One important assumption of the previous IICR studies was that the coalescence times obtained along the genome are sampled from the same distribution (under neutrality). Here this distribution will depend on class i. More precisely, the distribution of T_2^i is the exponential distribution with parameter $\mu_i = \frac{1}{\lambda_i}$, whose probability density function (pdf) is

$$f_i(t) = \mu_i e^{-\mu_i t}, i = 1 \dots K.$$

If the sampled loci are uniformly distributed along the genome and represent an unbiased sample of the genomic λ_i values, and if we denote by a_i the proportion of the genome corresponding to class i, we will then infer a genomic distribution of coalescence time with pdf

$$f(t) = \sum_{i=1}^{K} a_i f_i(t) = \sum_{i=1}^{K} a_i \mu_i e^{-\mu_i t}. \tag{1}$$

IICR expression and main properties

Denoting F the cumulative distribution function of T_2 and $f(t) = F'(t)$ its pdf, the IICR can be defined Mazet et al. (2016) in the most general case as:

$$\text{IICR}(t) = \frac{R(t)}{f(t)}$$

where

$$R(t) = \mathbb{P}(T_2 \geq t) = 1 - F(t).$$

For our model with K different λ_i, we then have from equation (1):

$$\text{IICR}(t) = -\frac{R(t)}{R'(t)} = \frac{\sum_{i=1}^{K} a_i e^{-\mu_i t}}{\sum_{i=1}^{K} a_i \mu_i e^{-\mu_i t}}. \tag{2}$$

Thus, the first result on the IICR is that, despite the panmixia and constant size of the population, it is straightforward to see that the IICR is not constant as soon as there are at least two different values of λ_i with non null proportion a_i across the genome. In other words, the assumption that there is more than one N_e in the genome means that the IICR will change with time. Classical interpretations of PSMC plots under panmixia will thus lead to the conclusion that the population size changed through time.

To be more specific on the nature of these changes, we now obtain the derivative of the IICR as a function of time (backward from present):

$$\mathrm{IICR}'(t) = \frac{R(t)R''(t) - R'(t)^2}{R'(t)^2}$$

which has the sign of

$$\begin{aligned}
R(t)R''(t) - R'(t)^2 &= \sum_{i=1}^{K} a_i e^{-\mu_i t} \sum_{j=1}^{K} a_j \mu_j^2 e^{-\mu_j t} - \sum_{i=1}^{K} a_i \mu_i e^{-\mu_i t} \sum_{j=1}^{K} a_j \mu_j e^{-\mu_j t} \\
&= \sum_{i=1}^{K} \sum_{j \neq i} a_i e^{-\mu_i t} a_j e^{-\mu_j t} \mu_j^2 - \sum_{i=1}^{K} \sum_{j \neq i} a_i e^{-\mu_i t} a_j e^{-\mu_j t} \mu_i \mu_j \\
&= \sum_{i=1}^{K} \sum_{j > i} a_i e^{-\mu_i t} a_j e^{-\mu_j t} (\mu_i^2 + \mu_j^2 - \mu_i \mu_j - \mu_j \mu_i) \\
&= \sum_{i=1}^{K} \sum_{j > i} a_i e^{-\mu_i t} a_j e^{-\mu_j t} (\mu_i - \mu_j)^2
\end{aligned}$$

This quantity is always positive so we can conclude that the IICR is *always increasing* from $t = 0$ to $t = +\infty$ (*i.e.* backward in time). In a stationary panmictic population, the fact that there are at least two N_e across the genome $(\lambda_i, i > 1)$ means that we will always infer a decreasing IICR (forward in time) typically interpreted as an apparent but spurious population size decrease. Interestingly, it could also be interpreted as the spurious presence of population structure since population structure can also generate similar changes in the IICR.

Detailed results with a two-class model

These properties can be observed in Figure 1 where we represent the simplest case with $K = 2$ classes of genomic regions. In this figure we present the IICRs for $\lambda_1 = 0.1$

and $\lambda_2 = 1$, for proportions of λ_2 (represented by the parameter a_2) varying from 0 to 1. Consistent with the choice made in most studies inferring past population size changes, time is plotted in log10 scale in this Figure and all others shown in the main text. Equivalent figures with time plotted in natural scale are provided in the Supplementary Material, because they may sometimes lead to slightly different interpretations.

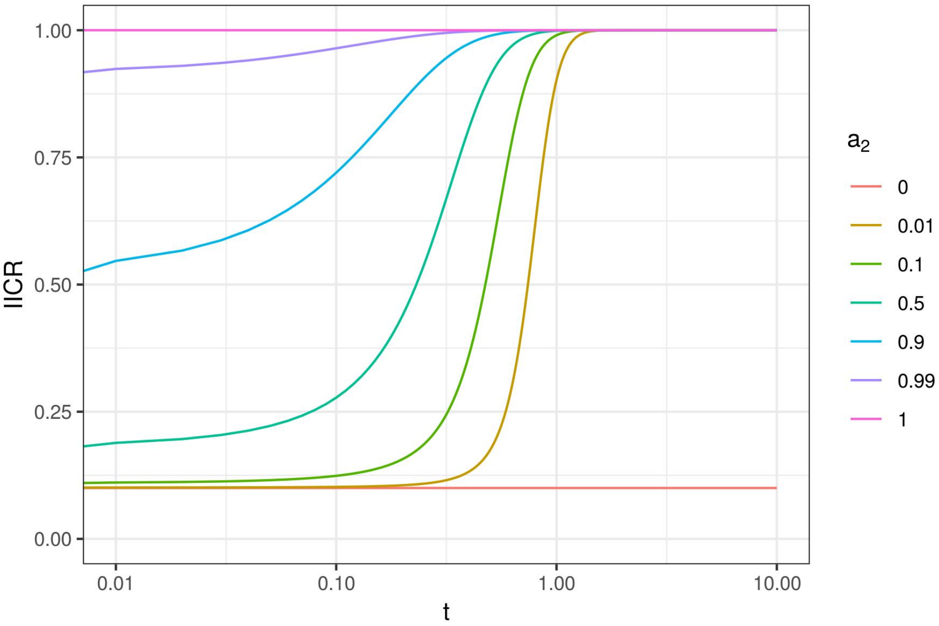

Figure 1: IICR curves for a panmictic model with $K = 2$ classes of genomic regions with constant size. Genomic regions of class i ($i = 1, 2$) have a constant population size $\lambda_i N$, with $\lambda_1 = 0.1$ and $\lambda_2 = 1$. Their frequencies are a_1 and a_2, respectively, with $a_1 + a_2 = 1$. The IICR curves are represented for a_2 values (representing neutrality, see main text) varying between zero and one. Time is plotted in log10 scale.

To simplify the interpretation of our results, we consider (by convention) throughout this manuscript that $\lambda_i = 1$ corresponds to the neutral regions of the genome, whether a_i, their relative proportion in the genome, is large or not. We thus do not necessarily consider that most of the genome is neutral in that sense. In this setting and in Figure 1, where $\lambda_1 = 0.1$ and $\lambda_2 = 1$, a_1 can be interpreted as the fraction of the genome showing reduced N_e by a multiplicative factor $\lambda_1 = 0.1$ as a consequence of positive or background selection.

Figure 1 thus shows that for small values of a_2 (*i.e.* when most of the genome is under N_e-reducing selection)the IICR is S-shaped, slowly increasing backward from $\lambda_1 = 0.1$ in the recent past to a plateau at $\lambda_2 = 1$ in the ancient past. For increasing a_2 values the IICR curves are becoming flatter as their left-most section flattens upward. When using a natural time scale (Figure S1) the curves simply seem to be shifted to the left for increasing a_2 values, the S shape being lost due to the truncation of the leftmost part of the curve. In other words, these curves start (in recent times) at increasing IICR values above $\lambda_1 = 0.1$ when the value of a_2 increases, but the curves always reach the same plateau at $\lambda_2 = 1$. However, and this is an important point, this plateau is reached earlier as a_2 increases. When a_2=1, only the plateau remains and the IICR is flat at $\lambda_2 = 1$ and when $a_2 = 0$, it is a flat at $\lambda_1 = 0.1$. Thus, when there is only one λ_i over the genome, the IICR is constant over time and equal to that value, as expected for a population with constant size $\lambda_i N$.

If we now assume that the only type of selection present in the genome increases the effective size by an order of magnitude, with a_1 and a_2 corresponding to $\lambda_1 = 1$ and $\lambda_2 = 10$, we obtain exactly the same figure with the only difference that it is rescaled (see Discussion and Figure S2). This figure now shows that even if most of the genome is neutral, tiny amounts of N_e-increasing selection strongly influence the IICR, as it always grows backward towards the plateau corresponding to the largest of the two λ_i values. It also shows that the recent IICR values can also be significantly greater than one, expected under neutrality.

Altogether Figures 1 and S2 suggest that there is a strong asymmetry between selection reducing (background and positive) or increasing (balancing) N_e in the genome in the way they affect IICR shapes. Balancing selection generates an ancient and high plateau at the level of λ_2, even for small proportions of a_2 (Figure S2), whereas positive and background selection generate a recent and relatively more modest decrease of the IICR for small values of a_1, even assuming, as in Figure 1, that these generate a ten-fold decrease in N_e (Figure 1).

Extension to more classes

While these observations are limited to the simplest case where there are only two λ_i values, it is straightforward to show that they hold for any number of classes. Indeed, the starting value of the IICR in a model with K λ_i values is

$$\text{IICR}(0) = \frac{1}{\sum_{i=1}^{K} a_i \mu_i} = \frac{1}{\sum_{i=1}^{K} \frac{a_i}{\lambda_i}} \tag{3}$$

and the limit value when $t \to +\infty$ is equal to

$$\frac{1}{\mu_{i_0}} = \lambda_{i_0} = \max_{i=1\ldots K}(\lambda_i).$$

The initial value IICR(0) is thus necessarily between the smallest and largest λ_i, as it is the harmonic mean of the λ_is weighted by their respective proportions a_i. The asymptotic value IICR($+\infty$) will always be the largest λ_i found in the genome, *independent* of its proportion. In other words, even if a minute proportion of the genome is under balancing selection, under panmixia the IICR should necessarily plateau in the ancient past to a value corresponding to that λ_i. One intuitive explanation for the IICR growing (backward in time) towards the largest λ_i is that the genes that are characterized by a large N_e

have much larger coalescence times than the rest of the genome. They thus contribute proportionately more to most ancient part of the IICR curve.

To further explore the influence of both types of selection (reducing and increasing N_e), we considered a model with 3 classes such that $\lambda_1 < 1$, $\lambda_2 = 1$ and $\lambda_3 > 1$ (Figure 2). In this Figure we set the three λ_i as $(\lambda_1, \lambda_2, \lambda_3) = (0.1, 1, 3)$. As above, $\lambda_1 < 1$ corresponds to genomic regions under linked positive or background selection, $\lambda_2 = 1$ corresponds to the neutral part of the genome and $\lambda_3 = 3$ to genomic regions under linked balancing selection. In the left panel, we considered a fixed small proportion of balancing selection ($a_3 = 0.01$), and allowed the proportions of neutral and positive or background selection to vary (a_1 varied from 0 to 0.8, and thus a_2 from 0.99 to 0.19). In the right panel, we considered a fixed and large proportion of positive or background selection ($a_1 = 0.5$) and varied the proportion of regions under balancing selection (a_3 from 0 to 10^{-1}), and thus the proportion of neutral regions too (a_2 between 0.5 and 0.4).

Figure 2 shows similarities with Figure 1. Specifically, both figures suggest that regions reducing N_e impact the IICR curves in the recent past whereas regions increasing N_e impact the IICR in the ancient past. This is worth stressing given that our model assumes that N_e is reduced (in class 1) or increased (in class 3) in a stationary way throughout the genealogical history of the sampled genes. Also, small proportions of balancing selection seem to generate much bigger changes than small proportions of positive or background selection, as shown by the comparison of the IICRs obtained for $a_1 = 0.01$ vs $a_1 = 0$ on one hand (left panel) and for $a_3 = 0.01$ vs $a_3 = 0$ on the other hand (right panel).

There are however differences between Figure 2 and Figure 1. The simple fact that we consider both N_e-reducing and N_e-increasing forms of selection and thus variable proportions of neutral genomic regions generates complex IICR curves, in which both forms of selection directly or indirectly impact the whole IICR curves. When neutral regions are frequent enough ($a_1 \leq 0.5$ and $a_3 \leq 0.01$), the IICR exhibits a plateau or a flattening at λ_2 in its middle section, but for larger values of either a_1 (left panel, $a_1 = 0.8$) or a_3 (right panel, $a_3 = 0.1$) the proportion of neutral genomic regions decreases and the IICR curve only exhibits a short inflexion corresponding to $\lambda_2 = 1$ before increasing backwards towards λ_3. An interesting pattern related to this intermediate plateau is observed on the left panel when a_3 is fixed:the IICR in the ancient past increases more and quicker (backward in time) for $a_1 = 0.8$ than for lower values of a_1, although a_1 models the proportion of low N_e regions in the region. This counterintuitive result likely comes from the fact that the proportion of neutral regions decreases when a_1 increases,so that the IICR directly increases to its highest possible value ($\lambda_3 = 3$).

Despite this complex interplay, Figure 2 provides some insights about our capacity to detect or quantify either type of selection based on the IICR. The left panel suggests that the proportion of the genome under positive or background selection can be assessed from this curve: for large values of a_1, there is a quick decline of the IICR (forward in time) followed by a low plateau around λ_1, whereas lower a_1 values see a more recent and gradual decrease of the IICR without any clear recent plateau. However, this distinction is far less visible when plotting on a natural scale (Figure S3), in which case a_1 values as different as 0.1 and 0.5 lead to quite similar IICRs. Besides, results on the importance of a_1 are likely exaggerated by the small value of λ_1 used in Figure 2, which implies a 10-fold reduction of N_e. In comparison, our choice of λ_3 only implies a 3-fold increase of N_e in Figure 2.

While the value of λ_3 (more generally of the highest λ_i) determines the plateau of the IICR, the proportion of this class (a_3) appears to determine to a large extent the speed of convergence (backward) to this ancient plateau (right panel). For the smallest a_3 values (0.1 or 0.01%), this ancient plateau is not reached within the figure (for $t \leq 10$) whereas a plateau corresponding to the neutral regions ($\lambda_2 = 1$) is observed for quite long periods. For the largest a_3 values considered here (1 or 10%), the convergence backward to the ancient plateau is so fast that the IICR does not exhibit the middle plateau around the neutral value, as already mentioned.

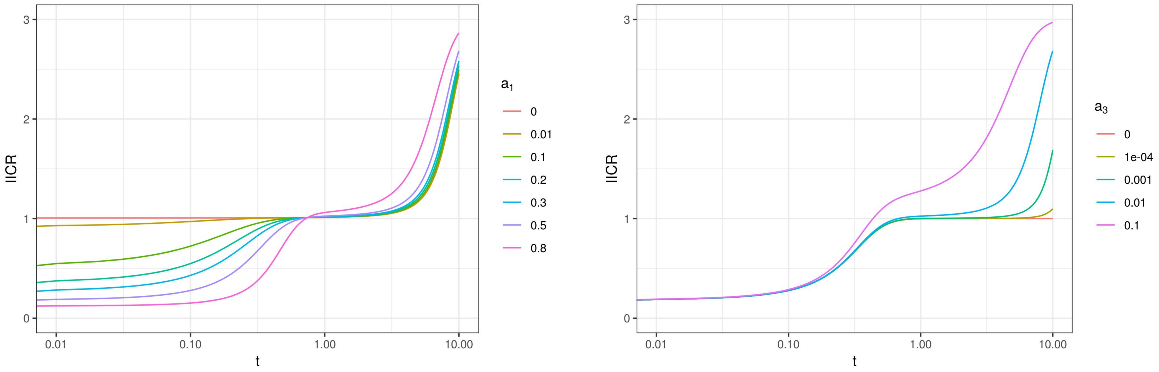

Figure 2: IICR for a panmictic model with $K = 3$ λ_i values such that $\lambda_1 < 1$, $\lambda_2 = 1$ and $\lambda_3 > 1$. The first class (or type) of genomic regions ($\lambda_1 < 1$) is meant to represent regions of the genome under (linked) positive or purifying selection and is modelled by a constant population size $\lambda_1 N$ with $\lambda_1 = 0.1$. Genomic regions of class 2 are meant to represent neutrality and they have a constant population size $\lambda_2 N$ where $\lambda_2 = 1$. Regions of class 3 are meant to represent genomic regions under (linked) balancing selection, they have a constant population size $\lambda_3 N$ with $\lambda_3 = 3$. Left panel: the frequency of class 3 is fixed at $a_3 = 0.01$ and the frequencies of classes 1 and 2 are allowed to vary. The frequency a_1 is given by the legend. Right panel: the frequency of class 1 is fixed at $a_1 = 0.5$ and the frequency of classes 2 and 3 are allowed to vary. The frequency a_3 is given by the legend.

While the value of λ_3 (more generally of the highest λ_i) determines the plateau of the IICR, the proportion of this class (a_3) appears to determine to a large extent the speed of convergence (backward) to this ancient plateau (right panel). For the smallest a_3 values

(0.1 or 0.01%), this ancient plateau is not reached within the figure (for $t \leq 10$) whereas a plateau corresponding to the neutral regions ($\lambda_2 = 1$) is observed for quite long periods. For the largest a_3 values considered here (1 or 10%), the convergence backward to the ancient plateau is so fast that the IICR does not exhibit the middle plateau around the neutral value, as already mentioned.

In any case, these results suggest that if selection can be seen as reducing or increasing N_e in a panmictic population, the strongest effect on the IICR seems to be disproportionately the result of the largest N_e, even though it may in practice affect ancient parts of the IICR curves that may not be easily reconstructed from real data. PSMC curves obtained from real data show a sharp increase (backward in time) in the very ancient past in several species, including humans and Neanderthals. While this ancient increase is usually ignored or interpreted as a statistical artefact resulting from the very low number of coalescence events dating back to this period, Figure 2 suggests that that it is possibly due to divergent alleles maintained by balancing selection. But it could also result from other factors of the ancient demographic history of species (Chikhi et al., 2018, Mazet et al., 2016). The interpretation of the IICRs represented in Figures 1 and 2 should indeed be done with caution given that the model used in the present section is panmictic whereas most species are likely to be structured. It is relevant to mention structure here because models of population structure also suggest a decrease of N_e (forward in time), visually similar to the model of selection considered here without structure. Models including both population structure and selection will be considered in a next section. Before coming to this we study a series of panmictic models inspired by research aiming at estimating variation in N_e in the genome of *Drosophila melanogaster* and *Homo sapiens*.

Towards realistic distributions of N_e

The above examples highlighted important and partly unexpected properties of the IICR when N_e is variable along the genome. However, they relied on arbitrary λ_i and a_i values. It is thus not clear to which extent they inform us on the impact of selection in real species. In this section we consider two model species for which variation in N_e has been documented or estimated, the fruit fly *Drosophila melanogaster* and humans (Figure 3).

In the case of *Drosophila melanogaster*, we compared two different distributions of λ_i over the genome. The first one was taken from the study of Elyashiv et al. (2016), who developed a method for inferring the distribution of fitness effects in different classes of functional annotations (UTRs, codons . . .) for both beneficial and deleterious mutations. This method requires polymorphism data from the focal species, divergence data with closely related species and precise recombination and annotation maps allowing to assess the selection constraints acting on each position in the genome. A by-product of their analysis is that an estimation of N_e can be obtained for sliding windows along the genome. Interestingly, these N_e values resulting from the strength of linked selection in each genomic region are defined as the inverse of the coalescence rate between two sequences and all computations rely on heterozygosity values observed between pairs of individuals. This suggests that the N_e estimates should be directly comparable with our λ_i values, which also correspond to the inverse of pairwise coalescence rates. The values of N_e estimated by Elyashiv et al. (2016) for 1Mb sliding windows in *Drosophila melanogaster* were downloaded at *https://github.com/sellalab/LinkedSelectionMaps*. Their distribution (top left panel) was converted into a discrete distribution of λ_i values with $K = 25$ classes using the *hist*() function of R. The IICR resulting from this distribution is shown in the top middle and right panels.

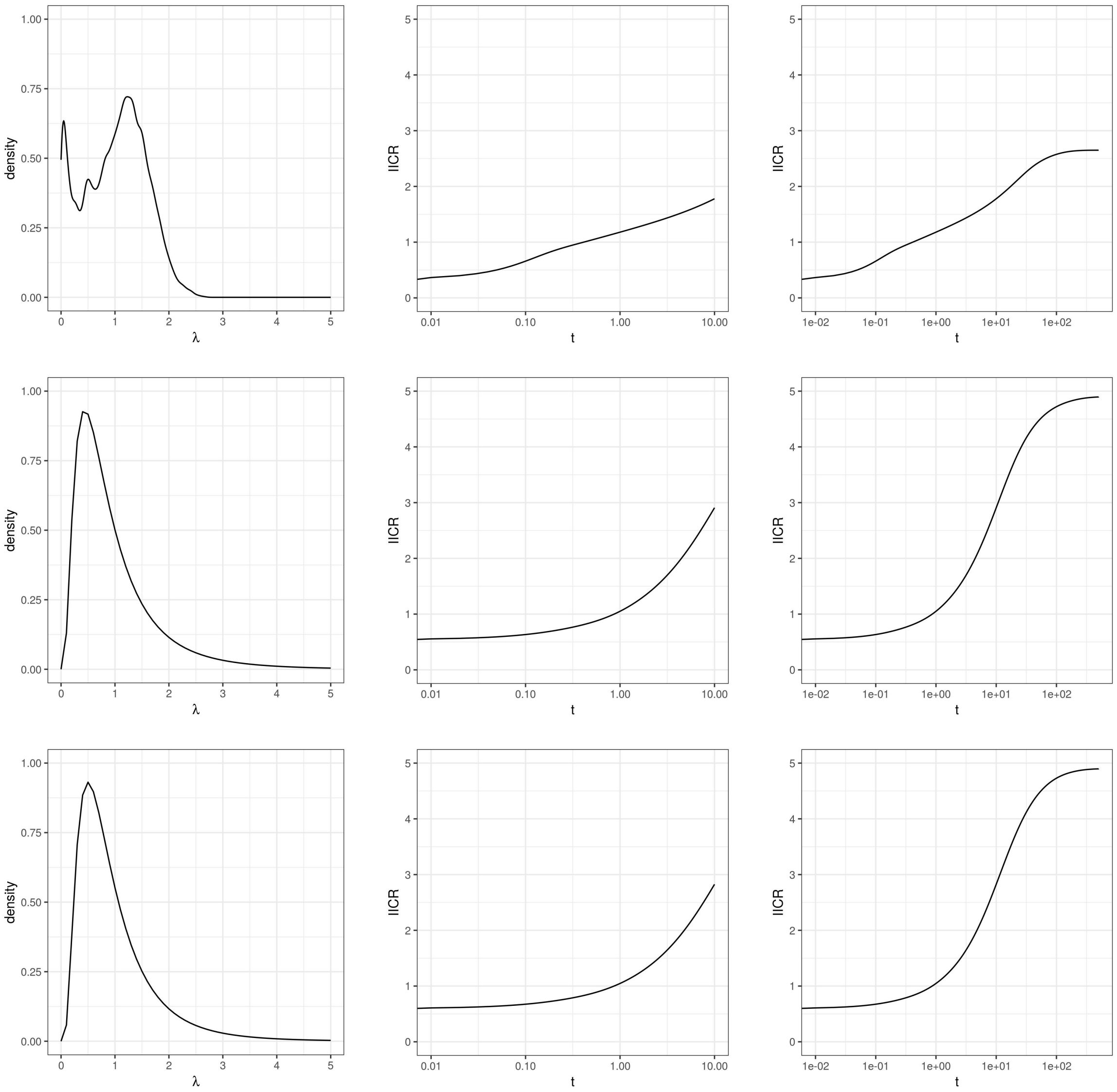

Figure 3: IICRs for panmictic models with large numbers of classes. This figure represents genome-wide distributions of λ_i (left panels) and the associated IICRs until $t = 10$ (middle panels) or $t = 500$ (right panels). Top panels: IICR for *Drosophila melanogaster* based on the N_e distribution estimated by Elyashiv et al. (2016). Middle panels: IICR for *D. melanogaster* based on the N_e distribution estimated by Gossmann et al. (2011) assuming a lognormal distribution. To make the two IICRs comparable, the distribution estimated by Elyashiv et al. (2016) (top left) was re-scaled to have an average of one, as assumed in the analysis of Gossmann et al. (2011) (middle left). Bottom panels: IICR for humans based on the N_e distribution estimated by Gossmann et al. (2011) assuming a lognormal distribution.

The second distribution used for this species was that estimated by Gossmann et al. (2011). While these authors also used polymorphism and divergence data, they focused on exons and did not aim at modelling the distribution of fitness effects. They assumed a log-normal distribution of N_e with mean value of 1 and estimated the scale parameter of this distribution from the observed data at several independent genes in the genome. Using the parameter obtained by this approach for *Drosophila melanogaster* and no recombination within genes (Table 1 of their study), we randomly sampled 100,000 values of N_e (or λ) under the log-normal distribution (middle left panel). A discrete distribution of the λ_i's and the associated IICR were then computed as explained above, filtering out large λ values (we arbitrarily excluded values above five). Indeed, it is not clear whether such large values would be realistic or statistical artifacts resulting from the use of a continuous distribution estimated mainly from smaller λ values. Also, they represent less than O.6% of the distribution. As a comparison with another species, we also applied this second approach with the scale parameter inferred by Gossmann et al. (2011) for humans (bottom panels).

Figure 3 shows that, when focusing on times from 0 to 10 (middle column), the three IICRs produced by these analyses have similar shapes. The value of the IICR at $t = 0$ is close to 0.5 for the three distributions. Interestingly, the two most similar IICRs were those based on the log-normal distribution estimated by Gossmann et al. (2011), despite the fact they correspond to two rather different species (fruit fly and humans).

The IICR resulting from the distribution of Elyashiv et al. (2016) for *Drosophila* differs from the other two on two aspects. First, it has a lower value at $t = 10$ (below two, whereas the others are close to three). One likely explanation is the smaller support of this distribution (up to $\lambda_i = 2.5$, versus $\lambda_i = 5$ for the others), which implies a smaller plateau of the IICR as demonstrated in previous section. This effect becomes clear when considering more ancient times (back to $t = 500$, right columns), but IICRs at such ancient times are unlikely to be observed from real data. Second, the IICR resulting from the distribution of Elyashiv et al. (2016) top middle panel) includes an inflexion between $t = 0.1$ and $t = 1$. This inflexion may be related to the mode observed for very low λ_i values with this approach (which probably results from the inclusion of regions with very low recombination where the impact of linked selection is substantial) but not with the approach of Gossmann et al. (2011). Indeed, a similar (although more pronounced) inflexion was observed for models combining selection both reducing and increasing N_e (Figure 2) but not for models with only one neutral and one selection class (Figure 1). Consistent with this hypothesis, filtering out λ values below 0.25 from the distribution of Elyashiv et al. (2016) lead to an IICR without inflexion (Figure S4).

Beyond these differences between the approaches of Gossmann et al. (2011) and Elyashiv et al. (2016), we note again that they resulted in rather similar IICRs, at least between $t = 0$ and $t = 10$. The magnitude of the decrease observed in these IICRs was also comparable to that expected from Figure 2 for small values of a_1 (e.g. $a_1 = 0.1$, top right panel). Consequently, a long term 5 fold IICR decrease (from $t = 10$ to $t = 0$ forward in time) could realistically be the result, in both humans and Drosophila, of a moderate proportion of loci with very small N_e (Figure 2, $a_1 = 0.1$, Figure 3, top) or from a larger proportion of loci with only slightly decreased N_e (Figure 3, middle and bottom), all as a consequence of linked selection.

Generalisation to more complex models

We can generalise equation (2) to more complex models by still assuming that the genome is divided into K groups of loci each characterized by a different coalescence rate history.

However, instead of describing this history by assuming panmixia and constant population size (λ_iN), we can study different demographic models with departures from these assumptions, including models with population structure, models with panmixia and population size changes, or models with transient changes in some of the λ_i values. In this more general framework, let us denote by $f_i(t)$ the *pdf* of the T_2 corresponding to the i-th class. If we keep the assumption that we can obtain a sufficiently large and unbiased number of independent loci across the genome, and if we denote by a_i the proportion of the genome described by f_i, then the IICR is:

$$\mathrm{IICR}(t) = \frac{\sum_{i=1}^{K} a_i R_i(t)}{\sum_{i=1}^{K} a_i f_i(t)}. \tag{4}$$

where $f_i(t) = -R_i'(t)$.

Models with population structure: the n-island model

One potential application of this general framework is to account for population structure when modelling each genomic class. To illustrate this idea, we first considered a model with $K = 2$, $\lambda_1 = 0.1$ and $\lambda_2 = 1$ as in Figure 1. Here we assumed that these two classes evolved under a n-island model with the same number of demes ($n = 10$), the difference in N_e being modelled through the use of different deme sizes in the two classes ($\lambda_1 N$ and $\lambda_2 N$) We further assumed that selection did not affect migration, so that the *per* generation migration rate m was the same for the two classes. In other words, selection reducing N_e is assumed to operate after migration and thus only affects coalescence rates, but not migration rates, of the two genomic regions. This implies that the scaled migration rate $M = 2Nm$ is identical in the two classes (time scale is still $2N$ here, but $\lambda_i N$ now refers to deme size rather than to the entire population size). One way of seeing this is by considering that there are $2N$ haploid genomes in each deme with scaled migration rate $2Nm$ and that selection acts on the different genomic regions by changing drift by a factor λ_i.

As already mentioned and exploited in previous studies on the IICR (Grusea et al., 2018, Mazet et al., 2016, Rodríguez et al., 2018), the distribution of coalescence times under a symmetrical n-island model can be derived analytically (Herbots, 1994). Extending these derivations to a model with general deme size $\lambda_i N$, instead of N in previous studies, we can show (see Appendix) that in this case we have

$$f_i(t) = p_i e^{-\alpha_i t} + (\frac{1}{\lambda_i} - p_i)e^{-\beta_i t} \tag{5}$$

with

$$\alpha_i = \frac{1}{2}\left(\frac{1}{\lambda_i} + n\gamma + \sqrt{\left(\frac{1}{\lambda_i} + n\gamma\right)^2 - \frac{4}{\lambda_i}\gamma}\right),$$

$$\beta_i = \frac{1}{2}\left(\frac{1}{\lambda_i} + n\gamma - \sqrt{\left(\frac{1}{\lambda_i} + n\gamma\right)^2 - \frac{4}{\lambda_i}\gamma}\right),$$

$$\gamma = \frac{M}{n-1}$$

and

$$p_i = \frac{\gamma - \alpha_i}{\lambda_i(\beta_i - \alpha_i)}.$$

By setting $\lambda_i = 1$ for all i we have the results of Mazet et al. (2016). The IICR of an n-island model with two classes of deme size can be obtained by computing $f_i(t)$ with each λ_i using Equation (5) and inserting the results into Equation (4).

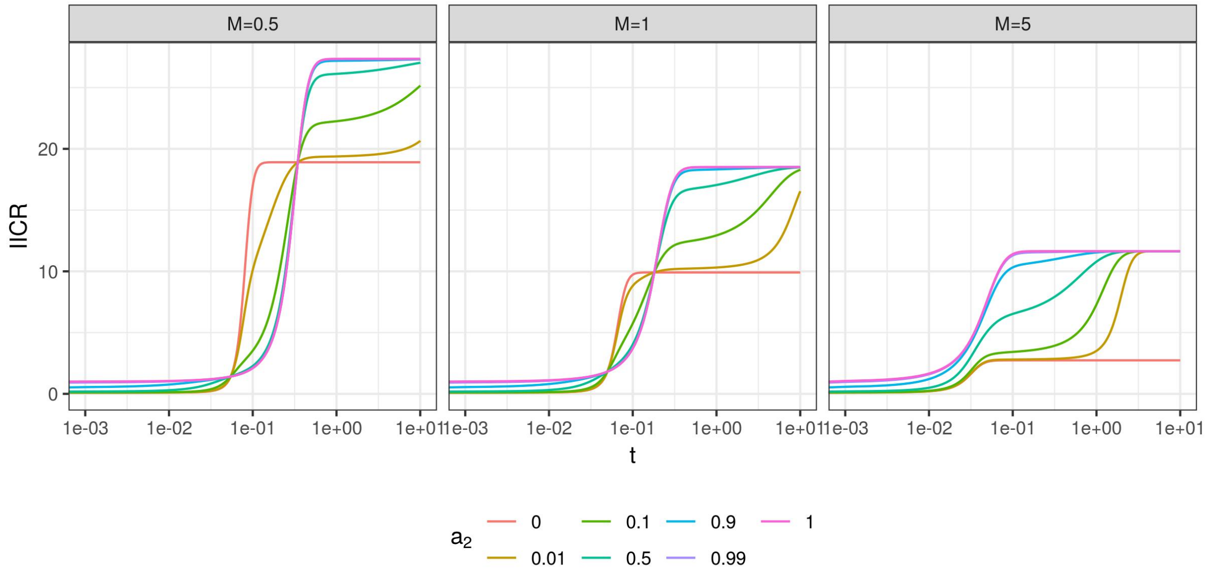

Figure 4: IICR curves for a symmetrical n-island model with $n = 10$ demes and $K = 2$ classes of genomic regions. Regions of class 1 and 2 have a constant deme size $2N\lambda_1$ and $2N\lambda_2$ with $\lambda_1 = 0.1$ and $\lambda_2 = 1$. Scaled migration rate $M = 4Nm$ is the same for the two classes, each panel corresponding to a different value of this parameter. Frequencies a_1 and a_2 of the 2 classes are given by the legend (having in mind that $a_1 + a_2 = 1$). For comparison with panmictic models (in particular those in Figure 1), time is scaled by the meta-population size $2Nn$ rather than by the deme size $2N$ as in Equation (5).

IICR curves obtained for this two class n-island model are shown in Figure 4 for different values of the scaled migration rate. For $M = 5$, they are similar to those shown in Figure 1. This was expected given that an n-island model with high migration ($M \gg 1$) should behave in a way that is similar to a panmictic model with population size $2Nn$, except in the recent past where the IICR of the n-island still reflects local deme size (Mazet et al., 2016). For lower migration rates, the two extreme models with $a_2 = 0$

(red curve) or $a_2 = 1$ (violet) show that a higher plateau of the IICR is observed as M decreases, which was again expected (Mazet et al., 2016).

For lower migration rates ($M \leq 1$ in Figure 4), models with rather large values of a_1 are hard to distinguish from the model with a_1=0 (no selection). For instance, the IICR with $a_2 = a_1 = 0.5$ is not very different from that with $a_2 = 1$, in contrast to Figure 1 where panmixia was assumed. This suggests that population structure may tend to mask the effect of positive or negative selection as long as a moderate part of the genome is under selection. On the other hand, the IICR with $a_2 = 0.01$ is more similar to that with $a_2 = 0$ than under panmixia. This suggests that, in the presence of population structure, models with pervasive selection (99% of the genome with $\lambda = 0.1$) may be interpreted as neutral models with small effective size (100% of the genome with $\lambda = 0.1$).

Another interesting observation from Figure 4 is the existence of a time window where the IICR is lower when a_2, corresponding to the largest N_e, is largest, *i.e.* the IICR is lower for models with a smaller part of their genome under selection reducing N_e. This time window occurs in the recent past and is wider for lower migration rates. This counter-intuitive result illustrates the limits of interpreting the IICR as a trajectory of effective size, as already outlined for several other demographic scenarios (Chikhi et al., 2018, Mazet et al., 2016). Outside this period, the IICR curves seem to always reach higher values when a_2 is larger. This is in particular the case for t close to 0, which is expected analytically (Equation (3)).

To check whether these conclusions may still hold for more realistic demographic scenarios, we next assume that each genomic class evolves under the non stationary n-island model estimated by Arredondo et al. (2021) to fit the observed PSMC of a modern human from Karitiana (Li and Durbin, 2011). This model includes 11 islands with symmetric migration and (diploid) deme size 1,380 and it assumes that these islands go through 4 changes of connectivity in the past: $M \approx 0.9$ ($m \approx 1.6e-4$) from present to 24,437 generations before present (BP), $M \approx 17.7$ ($m \approx 3.2e-3$) from 24,437 to 82,969 generations BP, $M \approx 2.5$ ($m \approx 4.5e-4$) from 82,969 to 107,338 generations BP, $M \approx 0.7$ ($m \approx 1.3e-4$) from 107,338 to 179,666 generations BP and $M \approx 1.1$ ($m \approx 2e-4$) in more ancient times. We define K classes of genomic regions: one neutral region with deme size N and $K-1$ other regions under selection with deme size $\lambda_i N$, for λ_i either smaller or larger than 1. Two different options are considered to model the heterogeneity of effective size along the genome: (i) the hypothetical three class model of Figure 2 with one class corresponding to positive or negative selection and one other corresponding to balancing selection (top panels), and (ii) the 25 class model of Figure 3 estimated from Gossmann et al. (2011)'s analysis of human real data (bottom panel).

We find that large values of a_1 could have a significant impact on the IICR in the period ranging from 10,000 to 30,000 generations ago (corresponding to 200-300,000 to 600-900,000 years ago). For instance with $a_1 = 0.8$, the IICR is around 17 in the most recent hump and around 5 in the most recent bump, versus 22 and 12 without selection (top left panel). However, this effect is very moderate when considering the λ_i distribution estimated by Gossmann et al. (2011) (bottom panel). Much more dramatic is the effect observed in the ancient past above 100,000 generations ($\approx$ 2-3 million years) before present, where the IICR with selection is significantly larger than the neutral IICR. This difference is driven by the part of the genome with large effective size (i.e. under balancing selection) and is found (with varying magnitude) in all scenarios.

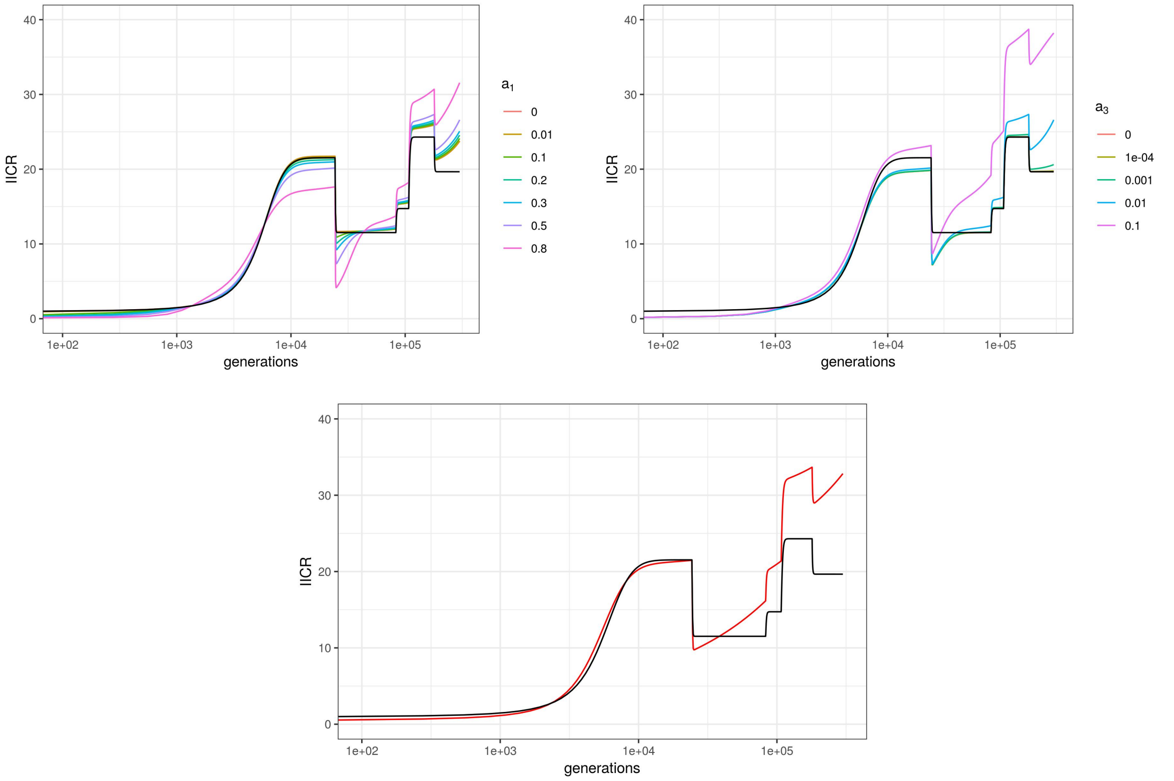

Figure 5: IICRs for demographic models combining population structure and linked selection in humans. The neutral part of the genome evolves under the non stationary n-island model estimated by Arredondo et al. (2021) to fit the observed PSMC of a modern human from Karitiana (Li and Durbin, 2011) **Good ref?**. This model includes 11 islands with (diploid) deme size $N = 1380$, whose connectivity varied along time according to a 3 step process (see the text for details). To account for selection, this neutral class only represents a fraction of the genome and other classes with lower or higher N_e are also considered. The number of these classes, their proportions and deme sizes (relative to the neutral class) are taken either from Figure 2 (top, where a_3 is fixed to 0.01 in the left panel, and a_1 fixed to 0.5 in the right one) or from Figure 3 (bottom, red line). The black curve on all panels depicts the IICR for this demographic scenario but without selection. Time is shown in generations and in log10 scale.

Panmictic models with transient selection

We finally apply this general framework to model the transient effect of recent selective sweeps, rather than the effect of recurrent positive, negative or balancing selection considered until now. For this analysis we consider a panmictic population. A similar question was tackled by Schrider et al. (2016), who showed in their Figure 5 the estimations obtained when applying the PSMC to a 15Mb genomic region that experienced one or several recent selective sweeps. We focus here on a scenario similar to theirs, with one single selective sweep and propose a model with different classes of λ_i that are time-dependent, which allows to approximate the resulting IICR. Although this model is built based on the expected variations of effective size (or coalescence rate) in a 15Mb region, we note that it also applies to a whole genome having experienced on average one recent selective sweep per 15 Mb region. In other words, our aim here is not to switch from the analysis of global to local IICRs, but rather to explore the local and implicitly global effects in a relatively realistic example.

To approximate the IICR resulting from a recent selective sweep, we assume that the effect of this sweep can be modelled by a reduction of effective population size that is limited both in time (from the emergence of the derived favorable allele to its eventual fixation in the population) and in "genomic space" (*i.e.* in a genomic neighborhood of this selected variant). More precisely, we consider that the region affected by the sweep on one side of the selected locus is of size

$$L = -\log(0.05)\frac{\alpha}{8Nr\log(\alpha)}$$

with N the diploid population size, r the per site recombination rate and $\alpha = 2Ns$ the scaled selection intensity (s being the fitness advantage of homozygotes carrying the selected mutation). This quantity corresponds to the distance in base pairs (bp) from the selected site such that heterozygosity is reduced by only 5% at the end of the sweep (Walsh and Lynch, 2018, chap. 8). To capture the fact that the reduction of effective size caused by the sweep depends on the physical distance to the selected site, we divide this affected region in 10 classes of same size $2\frac{L}{10}$ with increasing distance from the sweep, where the factor two results from the sweep extending on both sides of the selected site.

To quantify the reduction of effective size in a given class, we consider the genealogy at a neutral locus located d bp away from the selected site. This process can be modelled using a structured coalescent where lineages are either in the 'derived' or 'ancestral' background, depending on which allele at the selected locus they are associated with (to avoid any confusion, we remind here that this structure is a modelling facility and has nothing to do with the island structure considered in previous section. In this framework, ancestral recombination events creating or breaking the association with the derived allele can be seen as migration events from one background to the other (Kaplan et al., 1988). In the case of a complete selective sweep, lineages sampled at present all belong to the derived background, because the derived allele is then fixed in the population. Following previous studies on this topic, e.g. (Nielsen et al., 2005), we further assume a “star-like” model where these lineages can either (i) escape this derived background through recombination and stay in the ancestral background until the end of the sweep phase (i.e. at the time when the derived allele appeared, as we go backward in time) or (ii) coalesce all together at the end of the sweep phase. Actually, we slightly relax this second hypothesis and simply assume that their average coalescence time corresponds to the end of the sweep phase. The probability for each lineage to escape the sweep is approximately

$$q = 1 - e^{-4drN\log(\alpha)/\alpha}$$

Because lineages can only coalesce if they are in the same background (derived with probability $(1-q)^2$ or ancestral with probability q^2), we assume that the average coalescence rate during the sweep is

$$\mu_{sweep} = (1-q)^2\frac{1}{\tau} + q^2\frac{1}{2N}$$

where

$$\tau = 8N\log(\alpha)/\alpha$$

is the duration of the sweep (in generations). In this formula, $\frac{1}{\tau}$ approximates the average coalescence rate for two lineages not escaping the sweep, which follows from our assumption that the average coalescence time is τ, and $\frac{1}{2N}$ is the standard neutral coalescence rate which applies to two lineages having escaped the sweep.

In summary, the relative effective population size in a given genomic class affected by the sweep is equal to 1 before and after the sweep and to

$$\lambda_{sweep} = \frac{1/\mu_{sweep}}{2N}$$

during the τ generations of the sweep. A neutral class (with $\lambda = 1$ at all times) is also included to account for positions within the 15Mb segment but with physical distance to the selected site greater than L.

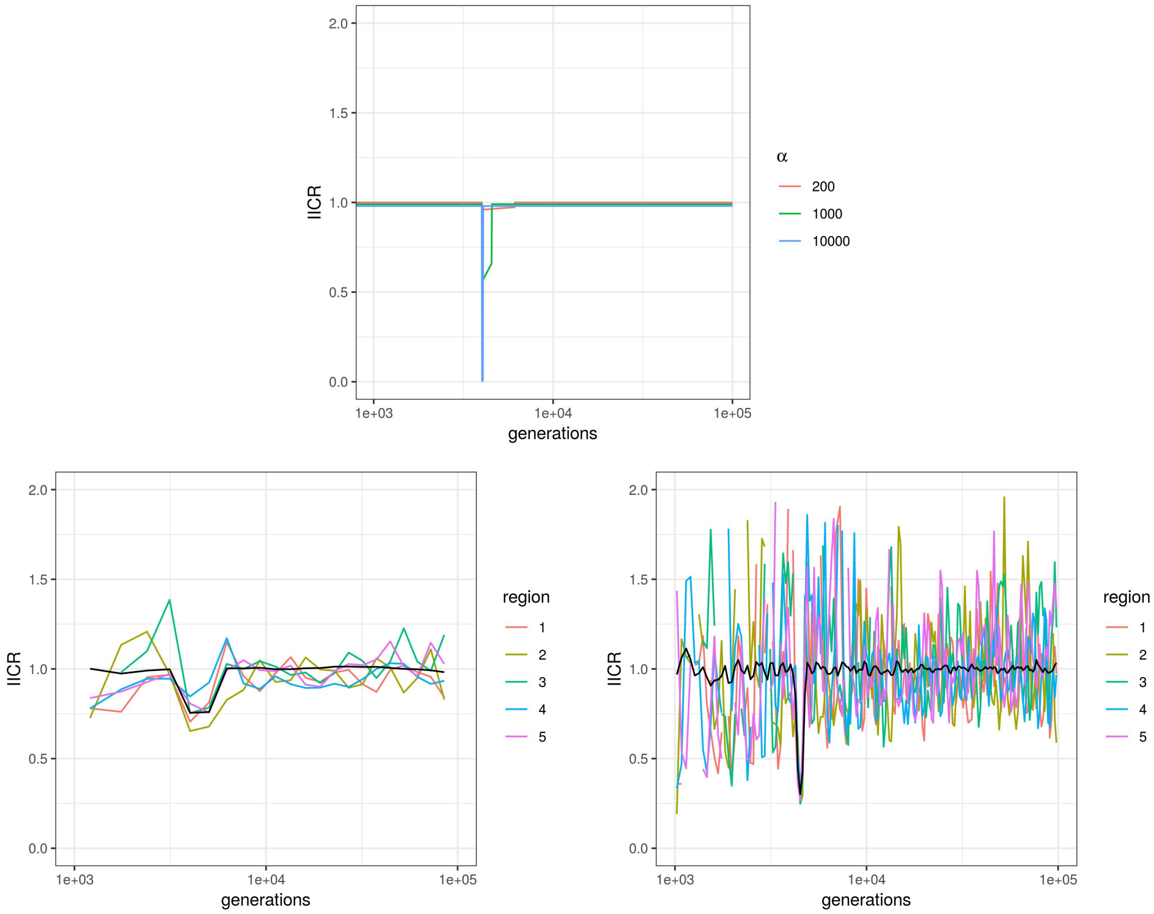

Figure 6: IICRs for a 15Mb region experiencing a single recent selective sweep. Parameter values were chosen to reproduce those in Figure 5 of Schrider et al. (2016): $N = 10000$ (diploid size), $r = 10^{-8}$ (per site recombination rate) and $t_0 = 4000$ generations before present (time where the derived allele got fixed). Times are given in generations and are shown in log10 scale. Top: Expected IICRs when modelling selection using a panmictic model with $K = 11$ classes of regions. Class 11 represents the neutral part of the region (unaffected by the sweep), with relative population size $\lambda_{11} = 1$. Class j $(1 \leq j \leq 10)$ represents a part of the region affected by the sweep, with a given physical distance from the selected site (which increases with j). Relative population size is equal to $\lambda_j = 1$ before and after the sweep and is decreased during the sweep to match the larger coalescence rate (see the text for more details). The proportion of each selected class $j \geq 10$ is $L/5$, where L is the size of the region affected by the sweep on either side of the selected site. Scaled selection intensity $\alpha = 2Ns$ was equal to 200, 1000 or 10000 (see the legend). Bottom: Empirical IICRs based on coalescence times simulated with the software *msms*, for $\alpha = 1000$. Two hundreds independent 15Mb regions were simulated. Colored lines show the IICRs for 5 of these regions (taken at random) and thus represent typical local IICRs. Black lines show the IICRs obtained when merging coalescence times from all regions, they thus correspond to genome-wide IICRs obtained for a 3Gb genome (200 × 15Mb) with one selective sweep every 15Mb. The number of time windows considered (i.e. of distinct estimated IICR values) was equal to 25 (left) or 200 (right) and the length of these windows was increasing exponentially backward in time, as in the PSMC approach.

As shown in Figure 6, top panel, the resulting IICR for $\alpha = 200$ (corresponding to $s = 0.01$ for $N = 10,000$) is very close to that of a neutral scenario. The IICR for $\alpha = 1000$ (corresponding to $s = 0.05$ for $N = 10,000$) shows a reduction of about one half at sweep time, similar to the average PSMC plot in Figure 6B of Schrider et al. (2016). The IICR for $\alpha = 10000$ (corresponding to $s = 0.5$ for $N = 10,000$ or to $s = 0.05$ for $N = 100,000$) shows a much stronger decline, down to almost zero. However, the IICR decline in our analysis is very localized in time, while the PSMC decline in (Schrider et al., 2016) extends for a longer period. Another important difference is that the PSMC plot in the simulations of Schrider et al. (2016) not only recovers the neutral value after the sweep but increases up to more than twice this value in the recent past. To understand these differences, we simulated coalescence times along a 15Mb region under the same sweep scenario, with $\alpha = 1000$, using the software *msms* (Ewing and Hermisson, 2010) and estimated the resulting empirical IICR as in Chikhi et al. (2018).

Similar to PSMC estimations, these empirical IICR estimations depend on the number of time windows considered, the assumption being that N_e is constant within each time window but may vary between time windows. In the bottom left panel of Figure 6, we consider 25 time windows, which corresponds to the order of magnitude used in most PSMC studies. The resulting IICR, averaged over 200 replicates, is transiently reduced around the sweep time and shows no increase above 1 in the recent past, similar to our theoretical prediction (top panel). However, the reduction of N_e is both longer and of lower magnitude than in our prediction, as in the PSMC plots of Schrider et al. (2016). In the bottom right panel, we consider 200 time windows and obtain an average IICR in which the magnitude and duration of the decrease is much more consistent with our theoretical prediction. IICRs from single replicates also correctly capture this reduction around the sweep time but are very noisy outside this period as a side effect of the finer time discretization. Altogether, these results show that modelling selective sweeps by local transient changes of population size leads to a reasonable approximation of the IICR (or equivalently of the genome-wide distribution of T_2) but that discretizing time using a limited number of time windows may lead to soften the true sweep signature by an averaging effect. The IICR increase following the sweep observed with the PSMC plots produced by Schrider et al. (2016) but not in our results (even when simulating sweeps) also outlines that the genome-wide distribution of T_2 may not be sufficient to characterize the genomic patterns left by selective sweeps in the data (some of them being exploited by PSMC). We come back to this point in the Discussion section.

Discussion

Effects of linked selection on the IICR

A now classical hypothesis in population genetics considers that linked selection can be modelled as a first approximation by a local change in effective population size (Hill and Robertson, 1966). Background selection and selective sweeps, which tend to reduce genetic diversity locally Charlesworth et al. (1993), Smith and Haigh (1974), are then seen as resulting in lower N_e values, whereas genomic regions under balancing selection are in contrast interpreted in terms of higher N_e values. In both cases, the impact of selection on genetic diversity or N_e is stronger for regions with lower recombination or higher selective constraints (number of selected sites, selection intensity) Charlesworth (2009). At the genome-wide level,linked selection appears thus to generate an apparent heterogeneity of N_e among genomic regions, reflecting the variations of the mode (increasing or decreasing N_e) and the intensity of linked selection (Gossmann et al., 2011, Jiménez-Mena et al., 2016a). Following this simplifying assumption, we described in this study the distribution of the coalescence time between two sequences (T_2) for models including variable classes of N_e along the genome. More precisely, we characterized the Inverse Instantaneous Coalescence Rate (IICR) (Mazet et al., 2016) of such models, a quantity that is equivalent to the T_2 distribution and corresponds to the output of the popular PSMC approach (Li and Durbin, 2011), which is generally (and in some cases at least wrongly) interpreted as the past temporal trajectory of N_e of the population or species under study. This analysis allowed us to predict the expected effects of linked selection on PSMC or related demographic inference approaches Schiffels and Durbin (2013).

One of the main conclusions of our work is that, under panmixia and constant population size, the existence of several classes of N_e (induced by linked selection) *always* results in a spurious signal of population size decline: the IICR of such models is a decreasing function (forward in time) whose highest value (reached in the ancient past) corresponds to the largest genomic N_e and lowest value (reached in the most recent past) to the harmonic mean of genomic N_e values weighted by their relative proportion in the genome (Figure 1, Equation 3). Specifically, we found that selection reducing N_e (background selection or sweeps) has a stronger effect on the IICR in the recent past, while selection increasing N_e (balancing selection) mainly influences the IICR in the intermediate and ancient past (Figure 2). There is a striking asymmetry between the two forms of selection: because the IICR plateau is determined by the class with the largest N_e independently of the proportion of this class, even a minute proportion of balancing selection can have a large effect on the IICR, whereas higher proportions of background selection or sweeps are necessary to generate significant and detectable effects on the IICR (Figure 2). Combining the two forms of selection by considering N_e distributions inferred from real data (Elyashiv et al., 2016, Gossmann et al., 2011) we found that linked selection is expected to cause a long term apparent five-fold decrease of the IICR in organisms such as humans or *Drosophila melanogaster* (Figure 3). However, we stress that these results assumed panmixia and constant population size.

Another important conclusion of our work is indeed that the effects of linked selection on the IICR mentioned above may be largely hidden by those of population structure. Considering a symmetrical n-island model, we observed for instance that even when a large proportion of the genome is influenced by selection reducing N_e the effect on the IICR could be difficult to see for models with reduced migration rates between islands (Figure 4). Focusing on humans we also considered a simple but reasonable demographic scenario of variable population structure Arredondo et al. (2021) together with a realistic genomic N_e distribution for this species Gossmann et al. (2011). We found that the largest and most visible effect of linked selection on the IICR was an ancient population size increase (backward in time) related to the presence of balancing selection (Figure 5, bottom).

The predominant influence of balancing selection outlined by our results may need to be toned down in more practical or realistic settings. Depending on the proportion of the genome under balancing selection, this effect may or may not be visible on real PSMC plots. For instance in humans, estimated PSMC plots extend to a few million years before present. Under the common assumptions of a diploid effective population size of 10,000 (under panmixia) and a generation time of about 20-30 years, one million years would correspond to a scaled time of five. Thus, human PSMC curves would fit within the panels represented in Figure 2, but the effect of balancing selection would not be easily detectable if $a_3 \leq 0.01$. For $a_3 \geq 0.01$, it might only be detected as a rapid backward increase in the most ancient parts of the PSMC curves, as also suggested when considering more realistic demographic and selection models (Figure 5). Such ancient increases are indeed observed in humans and a number of other species, but a further complication is that these patterns may also arise due to the low number of informative coalescence events available to PSMC in this ancient time period. PSMC analyses of genomic data simulated under realistic demographic scenarios, with and without balancing selection, will be necessary to investigate whether these ancient signatures of balancing selection can be disentangled from statistical artifacts. As a simple test we carried out additional simulations under an n-island model, generating genomic data under the demographic model of Figure 5 with a single genomic N_e (i.e. no selection). We applied PSMC to these data and found no ancient increase in the estimated trajectory compared to the expected IICR (Figure S5). These admittedly limited results suggest that the PSMC is not necessarily *statistically* biased in the ancient past, and that the signals observed in several species including humans and chimpanzees might be due to balancing selection or other forms of selection maintaining high levels of diversity over very long periods. One possible strategy to limit the influence of regions submitted to such forms of selection would be to first detect them and filter them out from the PSMC analysis. For the demographic scenario of Figure 5, we found that this would reduce the biases observed in the ancient past without affecting significantly other parts of the IICR (Figure S6).

The intriguing signature of background selection on the IICR

The framework developed in this study makes no particular distinction between positive and background selection, which are both modelled as leading to a reduction of N_e. Thus, one possible interpretation of our results would be that unaccounted background selection leads to a spurious signal of population decline. This conclusion is at odds with several previous studies, which concluded that unaccounted background selection may actually lead to a spurious signature of recent population expansion. For instance, Zeng and Charlesworth (2011) and Walczak et al. (2012) developed theoretical approximations of the genealogical process at a neutral locus linked to a site under negative selection and showed that this process shared many properties with that of an expanding population. The former study accounted for intra-locus recombination, whereas the latter ignored it. Several recent studies have applied demographic inference methods to genomic data simulated with and without linked background selection (Ewing and Jensen, 2016, Johri et al., 2021, Lapierre et al., 2016, Pouyet et al., 2018) and observed a signal of recent population expansion in the scenarios including selection. Finally, Johri et al. (2020) analyzed real data from an African population of *Drosophila melanogaster* with a new ABC demographic inference approach accounting for background selection. They estimated that the size of this population has been relatively constant for a few millions generations, while several previous studies on this or other related populations, which ignored background selection, estimated a strong recent population size increase.

Two main reasons may resolve this apparent paradox between these previous results and ours. First, we assume that linked selection can be modelled by a local change of N_e without any temporal dynamics (except in Figure 6 and related text), whose focus is specifically on recent selective sweeps). In particular, our results do not hold for demographic inference approaches based on the Site Frequency Spectrum (SFS), because weak background selection is expected to produce an excess of low frequency alleles, in particular singletons, which cannot be mimicked by just assuming a smaller N_e. Such an excess of rare alleles is also a classical signature of expanding populations, which may explain the conclusions of several of the studies mentioned above (Ewing and Jensen, 2016, Johri et al., 2020, Lapierre et al., 2016, Pouyet et al., 2018).

Second, even when focusing on pairwise statistics such as heterozygosity or T_2, the signature of population decline predicted by the IICR can only be observed if the data considered exhibit some heterogeneity in N_e. As it can easily be seen from Figure 1, panmictic models with either no ($a_2 = 1$) or only ($a_2 = 0$) selection do not show declining but constant IICRs. Consequently, a decline signature is not necessarily expected when analyzing a single locus under selection as in Zeng and Charlesworth (2011) or Walczak et al. (2012). It is also not necessarily expected when analyzing genome-wide data with homogeneous selective constraints along the genome. For instance, Johri et al. (2021) simulated genome-wide sequences including background selection by considering a regular alternance of functional (selected) and intergenic (neutral) regions of fixed and relatively small sizes: depending on the scenario, the size of a single 'unit' including one functional and one intergenic region ranged from $\approx$ 13 to 55 kb. The PSMC analyses of these sequences suggested a population under constant size or slight recent expansion. We believe that some of the results obtained by these (and possibly other) authors could be due to the fact that the data simulated with this approach do not exhibit enough heterogeneity in population sizes among (short) sliding windows over the genome. Such a regularity is at odds with observations made in different organisms (Elyashiv et al., 2016, Gossmann et al., 2011).

IICR predictions and PSMC estimations

Understanding the difference between our results and those of Johri et al. (2021) also leads to the fundamental question of the link between a PSMC curve and the IICR. As outlined in Mazet et al. (2016) and several subsequent studies, the IICR is a theoretical function associated to a given evolutionary model (and sampling scheme), while the PSMC is an estimation of this theoretical quantity. When population size history is homogeneous along the genome (i.e. $K = 1$ class), PSMC provides a very good estimation of the IICR (Mazet et al., 2016). But when population size history is heterogeneous along the genome, as considered here to approximate the effects of selection, the answer may depend on the scale (10kb? 100kb? 1Mb?) at which this heterogeneity is detectable. In other words, for a fixed proportion of genomic positions with reduced effective size due to linked selection, PSMC results may depend on the spatial clustering of these positions along the genome, while the IICR does not.

To explore this question, we tested whether genomic data including genome-wide heterogeneity of N_e at different scales could generate PSMC plots consistent with our IICR predictions. To do this we carried out a limited number of additional simulations in which, using the genomic sizes $\lambda_1 = 1$ and $\lambda_2 = 10$, we varied the lengths L_1 and L_2 of contiguous DNA chunks belonging to a given class, while keeping constant the proportions a_1 and $a_2 = 1 - a_1$ at which these classes are represented. We tested three values for the frequency a_1 (0.5, 0.9 and 0.99), and for each combination of these parameters we simulated two independent genomes of length 10^9 base pairs, where the two size classes were evenly spaced in the form:

$$\big(L_1, L_2, L_1, L_2, \ldots, L_1, L_2\big).$$

The lengths L_2 for the chunks of class 2 were chosen to be 10^6, 10^5 and 10^4 base pairs, and the lengths for the chunks of class 1 follow from the proportions a_1 and a_2. We found that PSMC estimations fit well our predictions for large chunks (10^6 and 10^5), but may highlight more complex and unpredicted patterns for smaller ones (Figure S7). This may explain the poor fit of our predictions with the PSMC results of Johri et al. (2021), where the heterogeneity of N_e was detectable only at very small scale ($\leq$ 55kb).

The recent selective sweep scenario considered in Figure 6 also nicely illustrates the potential difference between PSMC estimations and IICR predictions in the case of genomic heterogeneity. Simulating *genome sequences* in a single 15Mb region experiencing one recent selective sweep, Schrider et al. (2016) found that PSMC applied to these sequences would infer a bottleneck around the time of the sweep completion, generally followed by a more recent expansion exceeding the 'neutral' effective size. Simulating *coalescence times* under the same selective sweep scenario and estimating the IICR from these simulated values, we observed a similar bottleneck but no recent expansion. This difference likely results from the fact that short coalescence times are mostly clustered around the selected site in the real data, while for IICR estimation only their proportion over the 15Mb region matters. Including a transient change of N_e into our model with heterogeneous effective size along the genome, we can reproduce the main characteristics of the IICR with selection, but not completely of the PSMC with selection.

Overall, the results discussed in this section suggest that, although our study allows to describe the expected effects of linked selection on the overall distribution of coalescence times, specific simulations based on precise genomic annotations (positions and lengths of genes, local recombination rates ...) may be necessary to really assess potential PSMC biases in a given species.

Perspectives

The above discussion illustrates that the effects of linked selection on demographic inference are complex, as they not only depend on the type and intensity of selection but also on the inference approach applied and how data are summarized (SFS or T_2 based for instance) or the scale at which selection constraints vary along the genome. If the future confirms that linked selection is pervasive in the genome as claimed for several model species (Elyashiv et al., 2016, Pouyet et al., 2018) new demographic inference approaches accounting for linked selection and population structure will be needed. One way of achieving this objective is to jointly estimate demographic and selection parameters, as proposed in two recent studies relying on simulation based approaches, deep learning (Sheehan and Song, 2016) and Approximate Bayesian Computation (ABC) (Johri et al., 2020). These studies focused on relatively simple models, considering panmictic populations with a single population size change and only some types of selection (background selection in one study, sweeps and balancing selection in the other). To integrate more complex demographic scenarios, several recent studies considered interesting approaches, by assuming demographic models including two classes of N_e along the genome, one for neutral loci and one for loci under linked selection. The proportion of the two classes and the ratio of N_e between them were estimated together with other parameters of the demographic model, using either ABC (Rougemont and Bernatchez, 2018, Roux et al., 2016) or a modification (Rougemont et al., 2020, Rougeux et al., 2017) of the diffusion approach implemented in the software ∂a∂i (Gutenkunst et al., 2009). The models described in the present study propose another direction for the development of demographic inference methods by accounting for linked selection through variable classes of N_e along the genome, and using the IICR as summary statistic. An IICR-based inference framework was recently proposed for the estimation of non stationary n-island models and provided very encouraging results (Arredondo et al., 2021). Given the strong impact of linked selection on the IICR under panmixia, as described in the present study, we believe that a similar approach could allow to jointly infer parameters related to demographic history and to the N_e distribution. However, the results obtained under models of population structure suggest that it may be necessary to use the IICR in addition to other summaries of genomic diversity to overcome identifiability issues. Also, we should stress that separating the effects of population size change, selection and population structure is likely to be one of the major challenges of population genetics in the future.

Whether the objective is to predict potential effects of linked selection or to estimate linked selection parameters from real data, two nice features, and possible advantages of an IICR-based approach such as the one considered here are flexibility and speed of computation. Our approach allows to simultaneously include different forms of selection and to combine linked selection with arbitrary demographic models. The examples considered here included stationary panmictic and n-island models (Figure 2 and 4) and non stationary island models (Figure 5). We also considered different distributions of λ_i and temporal variation of λ_i. Our approach can easily be extended to more general structured models including temporal population size variations. In the case of structured models, variable migration rates along the genome may be considered, to mimic variation in introgression rates. Indeed, we could either decreasing M in the linked selection class to account for possible effects of selection on migration success or introduce new classes with lower M values in order to model possible barriers to gene flow (Roux et al., 2016). As outlined in Figure 6, transient selection can be modelled by including population size changes in one class, and this approach could also be extended to model more complex fluctuating selection effects.

Whatever the complexity of the model considered, the associated IICR can be computed exactly in a very small time using the rate matrix approach described in (Rodríguez et al., 2018) or Arredondo et al. (2021), which allows to efficiently explore a very large number of scenarios or parameter values. As previously mentioned, the main limitation of the IICR approach described in this study is that it focuses on pairs of sequences. It provides information that is complementary to that provided by the SFS, as we have noted elsewhere Arredondo et al. (2021), Chikhi et al. (2018) This means that some effects of weak background selection or selective sweeps may be visible on the SFS but not on the IICR. Currently we have mainly focused on the IICR as defined for a pair of sequences, but extensions to multiple sequences might provide additional information on the distribution of higher order coalescence times (T_3, T_4, ...)., hence allowing a finer characterization of selective and neutral processes.

To conclude we have used the IICR as a way to explore important ideas that are central to population genetics such as the notion of effective size (see also Chikhi et al. (2018), Mazet et al. (2016) for discussions on these questions), drift and selection. We wished to re-open discussions regarding the influence of selective and neutral processes on genetic diversity, some of them general and theoretical, others more specific and practical: Can selection be modelled as a genomic variation in N_e? What are the limits of such an approximation? How robust is it? Does an IICR perspective provide interesting outcomes? Can N_e variation along the genome be detected in real genomes by applying the PSMC method of (Li and Durbin, 2011) or related approaches? Can we infer the presence of linked selection in humans from the PSMC plots or SFS histograms observed in this and other species? These are exciting questions to ask and the recent years have shown that they are at the heart of modern population genetics.

Data availability statement

Code used to generate the exact and simulated IICRs shown in this study can be found at https://github.com/sboitard/IICR_selection.

Acknowledgements

Armando Arredondo was funded by the Université Fédérale Toulouse Midi Pyrénées (UFTMiP) and the Région Occitanie (formerly Midi-Pyrénées) with PhD grant No. 31I2017M248. Lounès Chikhi was funded by Fundação para a Ciência e Tecnologia (ref.

PTDC-BIA-EVL/30815/2017). Olivier Mazet and Lounès Chikhi were funded by the 2015–2016 BiodivERsA COFUND call for research proposals, with the national funders ANR (ANR-16-EBI3-0014) and the Fundação para a Ciência e Tecnologia ref. Biodiversa/0003/2015 and PT-DLR (01LC1617A). This work was also supported by the LABEX entitled TULIP (ANR-10-LABX-41 and ANR-11-IDEX-0002-02) as well as the LIA BEEG-B (Laboratoire International Associé-Bioinformatics, Ecology, Evolution, Genomics and Behaviour). We acknowledge an Investissement d'Avenir grant of the Agence Nationale de la Recherche (CEBA: ANR-10-LABX-25-01).

Supplementary Material

Derivation of the pdf of T_2 in a n-island model

We derive here the pdf density of T_2, the coalescence time of two lineages sampled in the same deme (resp. different deme), in an n-island model. We follow the identity by descent approach used in Durrett's process (Durrett, 2008, p. 150). The size of each deme is λN, the probability of each lineage to migrate from a deme to another each generation is m, and the per locus mutation rate is u. Define the rescaled mutation and migration rates by $\theta = 4Nu$ and $M = 4Nm$. Note that two lineages coalesce at rate $c = \frac{1}{\lambda}$ when they are in the same deme, migrate at rate $2m.2N = M$ and experience mutations at rate $2u.2N = \theta$.

Let $p_s(\theta)$ and $p_d(\theta)$ be the probabilities that two lineages are identical by descent when they are chosen in the same or different demes. Following back two lineages from the same deme, three different events can occur: a coalescence with probability $\frac{c}{c+\theta+M}$, a migration with probability $\frac{M}{c+\theta+M}$ and a mutation with probability $\frac{\theta}{c+\theta+M}$. If lineages are in different demes, the only possible events are mutation, with probability $\frac{\theta}{\theta+M}$ and migration. In this second case lineages arrive in the same deme with probability $\frac{1}{n-1}$ and stay in different ones with probability $\frac{n-2}{n-1}$. Hence we have the two coupled equations:

$$p_s(\theta) = \frac{c}{c+M+\theta}.1 + \frac{M}{c+M+\theta}.p_d(\theta),$$

and

$$p_d(\theta) = \frac{M/(n-1)}{M+\theta}.p_s(\theta) + \frac{M(n-2)/(n-1)}{M+\theta}.p_d(\theta).$$

The second equation gives

$$\left(1 - \frac{M(n-2)}{(n-1)(M+\theta)}\right) p_d(\theta) = \frac{M}{(n-1)(M+\theta)} p_s(\theta)$$
$$\Leftrightarrow \frac{\theta(n-1)+M}{(n-1)(M+\theta)} p_d(\theta) = \frac{M}{(n-1)(M+\theta)} p_s(\theta)$$
$$\Leftrightarrow p_d(\theta) = \frac{M}{\theta(n-1)+M} p_s(\theta).$$

We then inject in the first equation:

$$p_s(\theta) = \frac{c}{c+M+\theta} + \frac{M}{c+M+\theta}\frac{M}{\theta(n-1)+M} p_s(\theta)$$

hence

$$p_s\left(1 - \frac{M^2}{(c+M+\theta)(\theta(n-1)+M)}\right) = \frac{c}{c+M+\theta}$$

and since

$$(c+M+\theta)(\theta(n-1)+M) - M^2 = \theta^2(n-1) + \theta(c(n-1)+Mn) + cM,$$

we get

$$p_s = \frac{c(\theta(n-1)+M)}{\theta^2(n-1)+\theta(c(n-1)+Mn)+cM} = \frac{c(\theta+\gamma)}{\theta^2+\theta(c+n\gamma)+c\gamma}$$

and

$$p_d = \frac{cM}{\theta^2(n-1)+\theta(c(n-1)+Mn)+cM} = \frac{c\gamma}{\theta^2+\theta(c+n\gamma)+c\gamma}$$

with

$$\gamma = \frac{M}{n-1}.$$

Let's now note that the probability $p_s(\theta)$ that two lineages has reached their common ancestor without undergoing any mutation is also the expected value $\mathbb{E}\left(e^{\theta T_2}\right)$. In other words, p_s is the Laplace transform of T_2. It can be inverted by looking for the roots of $\theta^2+\theta(c+n\gamma)+c\gamma$. Let $\Delta = (c+n\gamma)^2-4c\gamma$, then

$$p_s(\theta) = \frac{c(\theta+\gamma)}{(\theta+\alpha)(\theta+\beta)} = \frac{a}{\theta+\alpha}+\frac{b}{\theta+\beta}$$

with

$$\alpha = \frac{1}{2}\left(c+n\gamma+\sqrt{\Delta}\right),$$

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

$$a = \frac{c(\gamma-\alpha)}{\beta-\alpha}$$

and

$$b = \frac{c(\gamma - \beta)}{\alpha - \beta} = c - a.$$

Hence the probability density function of T_2 is:

$$f_{T_2}(t) = ae^{-\alpha t} + (c - a)e^{-\beta t}.$$

Note that $-\alpha$ and $-\beta$ are the non zero eigenvalues of the Q-matrix, $-\beta$ being the closest to 0, and we have the relationships $\alpha + \beta = c + n\gamma$ and $\alpha\beta = c\gamma$. Note also that we could similarly obtain the pdf distribution of the coalescence time of two lineages sampled in different demes, as p_d is its Laplace transform as well.

873 Supplementary figures

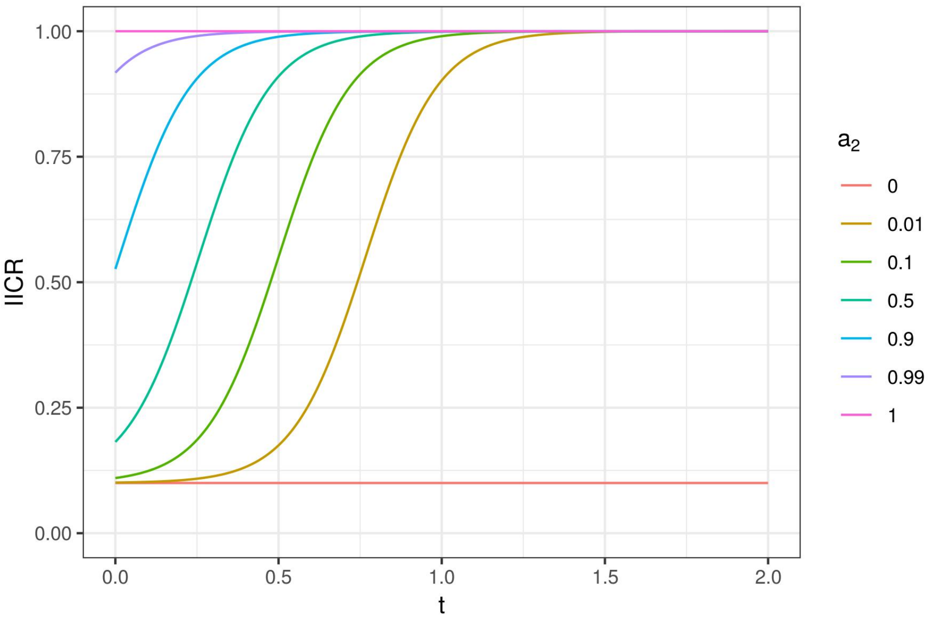

Figure S1: IICR curves for a panmictic model with $K = 2$ classes of genomic regions with constant size. Same as Figure 1 except that time is plotted in natural scale.

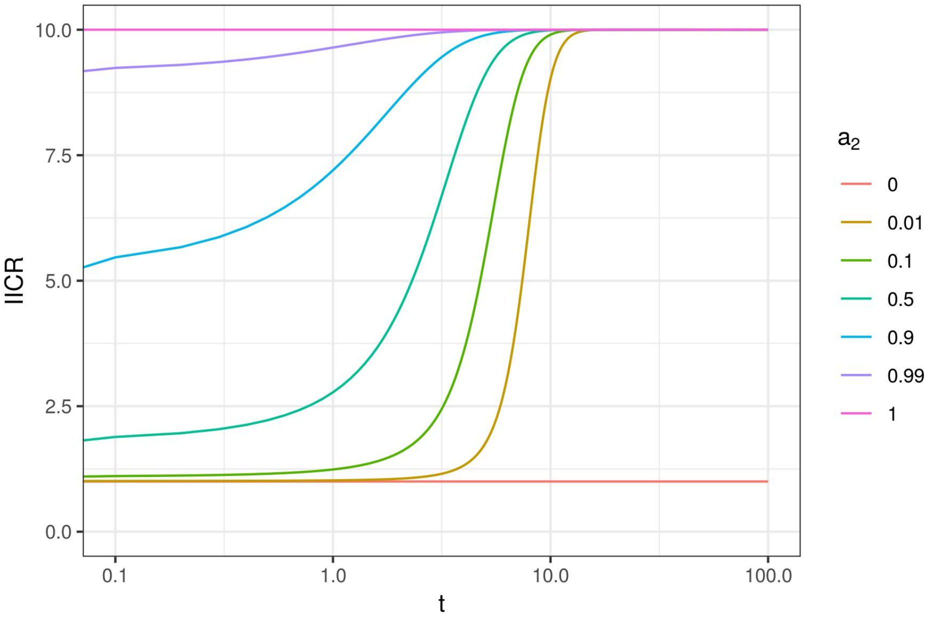

Figure S2: IICR curves for a panmictic model with $K = 2$ classes of genomic regions with constant size. Same as Figure 1 with $\lambda_1 = 1$, $\lambda_2 = 10$ and time from 0 to 100 (in log10 scale)

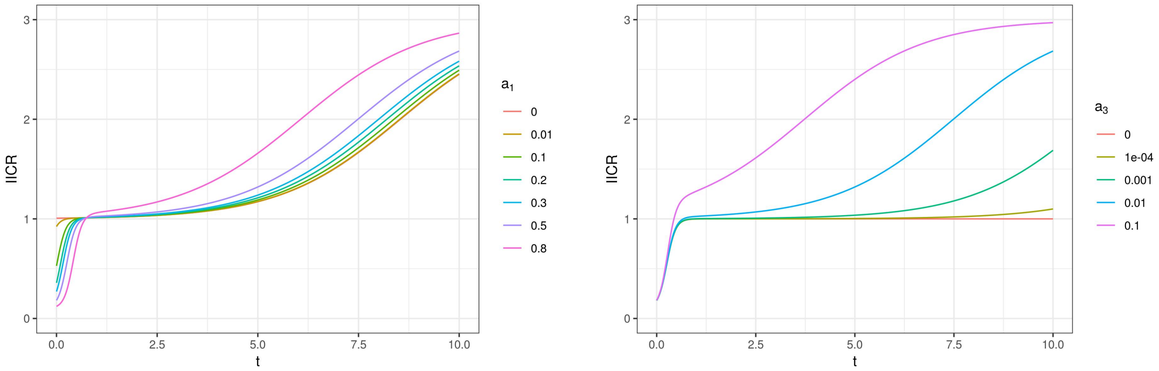

Figure S3: IICR for a panmictic model with $K = 3$ λ_i values such that $\lambda_1 < 1$, $\lambda_2 = 1$ and $\lambda_3 > 1$. Same as Figure 2 except that time is plotted in natural scale.

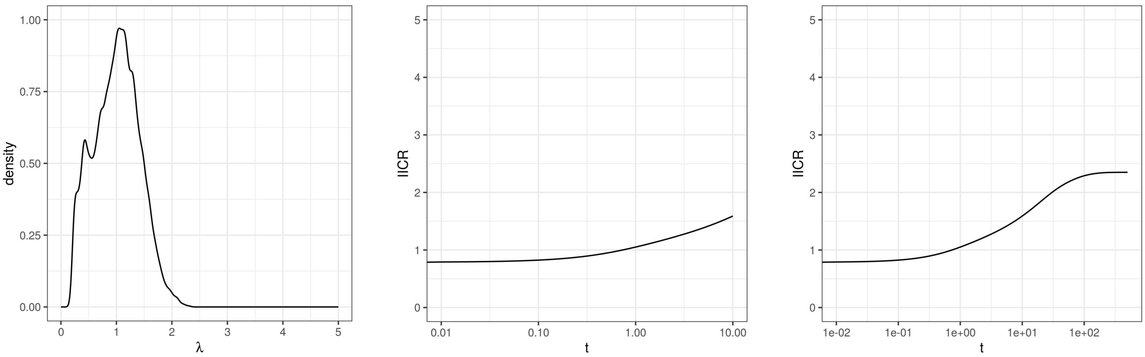

Figure S4: IICR obtained when removing low N_e values from the distribution estimated by Elyashiv et al. (2016). This truncated distribution (rescaled to have a mean of 1 as the others) is shown on the left panel. The associated IICR is shown until $t = 10$ (middle panel) or $t = 500$ (right panel), in log10 scale.

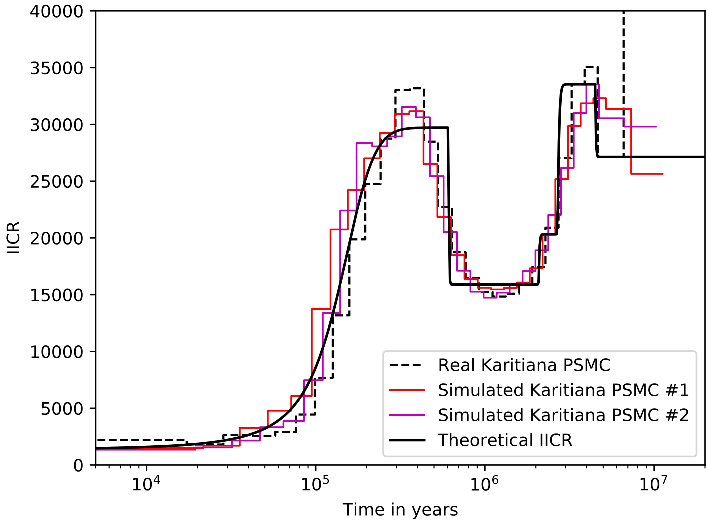

Figure S5: PSMC curves of simulated data under a non-stationary n-island model. We show in black the exact IICR corresponding to an inferred n-island model for a Karitiana individual in Arredondo et al. (2021). In color, we show various PSMC curves obtained by independently simulating genomic sequences under this structured model. The real PSMC curve for this Karitiana individual is represented by the dashed plot (Prado-Martinez et al., 2013). The horizontal axis is scaled in years, with a generation time of 25 years.

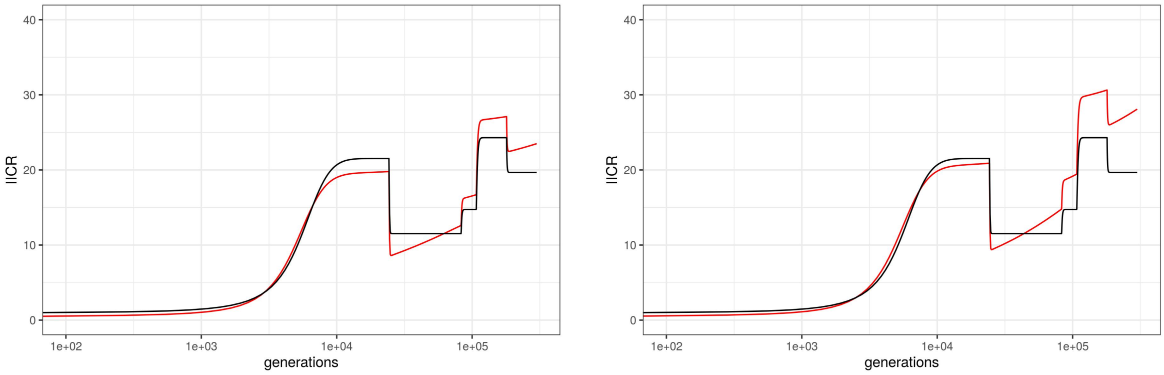

Figure S6: IICRs for demographic models combining population structure and linked selection in humans. Same as Figure 5, bottom panel, except that λ values greater than 2 (left) or 3 (right) were filtered out from the distribution in order to mimic a situation were loci under balancing selection could be detected and removed before computing the IICR. The resulting truncated distribution was rescaled.

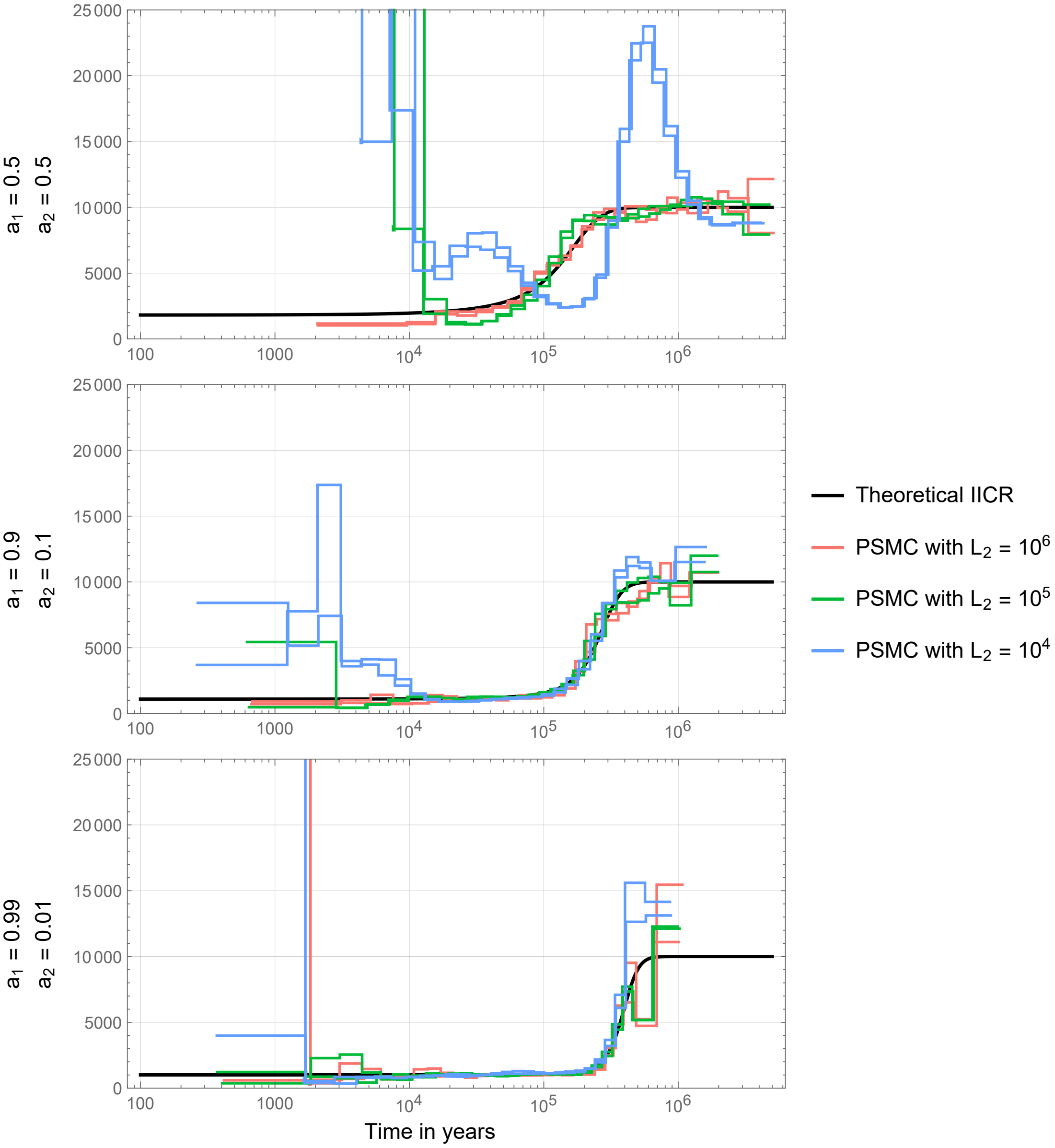

Figure S7: Comparison between theoretical IICR and inferred PSMC. For each frequency distribution (a_1, a_2) of the two size classes $\lambda_1 = 1$ and $\lambda_2 = 10$ we show the corresponding theoretical IICR (black) and two independent PSMC simulations for three variations of the chunk length L_2. In each case, $L_1 = \frac{a_1}{a_2} L_2$; the simulated sequence has a total length of 10^9 base pairs and the two class chunks are evenly alternated in the form $(L_1, L_2, L_1, \dots, L_2)$. The effective size N_1 was chosen as 1000. The horizontal axis is scaled in years, with a generation time of 25 years.

References

Arredondo, A., Mourato, B., Nguyen, K., Boitard, S., Valcarce, W. R. R., Noûs, C., Mazet, O., and Chikhi, L. (2021). Inferring number of populations and changes in connectivity under the n-island model. *Heredity*.

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10(3):195.

Charlesworth, B. (2010). *Elements of evolutionary genetics*. Roberts Publishers.

Charlesworth, B., Morgan, M., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–1303.

Chikhi, L., Rodriguez, W., Grusea, S., Santos, P., Boitard, S., and 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.

Comeron, J. M. (2017). Background selection as null hypothesis in population genomics: insights and challenges from drosophila studies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1736):20160471.

Durrett, R. (2008). *Probability models for DNA sequence evolution*. Springer.

Elyashiv, E., Sattath, S., Hu, T. T., Strutsovsky, A., McVicker, G., Andolfatto, P., Coop, G., and Sella, G. (2016). A genomic map of the effects of linked selection in drosophila. *PLoS genetics*, 12(8):e1006130.

Ewing, G. and Hermisson, J. (2010). Msms: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16):2064–2065.

Ewing, G. B. and Jensen, J. D. (2016). The consequences of not accounting for background selection in demographic inference. *Molecular ecology*, 25(1):135–141.

Gossmann, T. I., Woolfit, M., and Eyre-Walker, A. (2011). Quantifying the variation in the effective population size within a genome. *Genetics*, 189(4):1389–1402.

Grusea, S., Rodriguez, W., Boitard, S., Chikhi, L., and Mazet, O. (2018). Coalescence times for three genes are sufficient to detect population structure. *Journal of Mathematical Biology*, xxx(x):xxx–xxx.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5(10):e1000695.

Herbots, H. M. J. D. (1994). *Stochastic models in population genetics: genealogy and genetic differentiation in structured populations*. PhD thesis.

Hill, W. G. and Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetical Research*, 8(3):269–294.

Jensen, J. D., Payseur, B. A., Stephan, W., Aquadro, C. F., Lynch, M., Charlesworth, D., and Charlesworth, B. (2019). The importance of the neutral theory in 1968 and 50 years on: a response to kern and hahn 2018. *Evolution*, 73(1):111–114.

Jiménez-Mena, B., Bataillon, T., et al. (2016a). Heterogeneity in effective population size and its implications in conservation genetics and animal breeding. *Conservation genetics resources*, 8(1):35–41.

Jiménez-Mena, B., Tataru, P., Brøndum, R. F., Sahana, G., Guldbrandtsen, B., and

Bataillon, T. (2016b). One size fits all? direct evidence for the heterogeneity of genetic drift throughout the genome. *Biology letters*, 12(7):20160426.

Johri, P., Charlesworth, B., and Jensen, J. D. (2020). Toward an evolutionarily appropriate null model: Jointly inferring demography and purifying selection. *Genetics*, 215(1):173–192.

Johri, P., Riall, K., Becher, H., Excoffier, L., Charlesworth, B., and Jensen, J. D. (2021). The impact of purifying and background selection on the inference of population history: problems and prospects. *Molecular Biology and Evolution*. msab050.

Kaplan, N. L., Darden, T., and Hudson, R. R. (1988). The coalescent process in models with selection. *Genetics*, 120(3):819–829.

Kern, A. D. and Hahn, M. W. (2018). The neutral theory in light of natural selection. *Molecular biology and evolution*, 35(6):1366–1371.

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press.

Lapierre, M., Blin, C., Lambert, A., Achaz, G., and Rocha, E. P. (2016). The impact of selection, gene conversion, and biased sampling on the assessment of microbial demography. *Molecular biology and evolution*, 33(7):1711–1725.

Lewontin, R. C. (1974). *The genetic basis of evolutionary change*, volume 560. Columbia University Press New York.

Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496.

Mazet, O., Rodriguez, W., Grusea, S., Boitard, S., and Chikhi, L. (2016). On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference. *Heredity*, 116(4):362–371.

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., and Bustamante, C. (2005). Genomic scans for selective sweeps using snp data. *Genome research*, 15(11):1566–1575.

Ohta, T. (1992). The nearly neutral theory of molecular evolution. *Annual review of ecology and systematics*, 23(1):263–286.

Pouyet, F., Aeschbacher, S., Thiéry, A., and Excoffier, L. (2018). Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *Elife*, 7:e36317.

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O'Connor, T. D., Santpere, G., et al. (2013). Great ape genetic diversity and population history. *Nature*, 499(7459):471–475.

Rodríguez, W., Mazet, O., Grusea, S., Arredondo, A., Corujo, J. M., Boitard, S., and Chikhi, L. (2018). The iicr and the non-stationary structured coalescent: towards demographic inference with arbitrary changes in population structure. *Heredity*, 121(6):663.

Rougemont, Q. and Bernatchez, L. (2018). The demographic history of atlantic salmon (salmo salar) across its distribution range reconstructed from approximate bayesian computations. *Evolution*, 72(6):1261–1277.

Rougemont, Q., Moore, J.-S., Leroy, T., Normandeau, E., Rondeau, E. B., Withler, R. E., Van Doornik, D. M., Crane, P. A., Naish, K. A., Garza, J. C., et al. (2020).

Demographic history shaped geographical patterns of deleterious mutation load in a broadly distributed pacific salmon. *PLoS genetics*, 16(8):e1008348.

Rougeux, C., Bernatchez, L., and Gagnaire, P.-A. (2017). Modeling the multiple facets of speciation-with-gene-flow toward inferring the divergence history of lake whitefish species pairs (coregonus clupeaformis). *Genome biology and evolution*, 9(8):2057–2074.

Roux, C., Fraïsse, C., Romiguier, J., Anciaux, Y., Galtier, N., and Bierne, N. (2016). Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLOS Biology*, 14(12):1–22.

Schiffels, S. and Durbin, R. (2013). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 8(46):919–925.

Schrider, D. R., Shanku, A. G., and Kern, A. D. (2016). Effects of linked selective sweeps on demographic inference and model selection. *Genetics*, 204(3):1207–1223.

Sheehan, S. and Song, Y. S. (2016). Deep learning for population genetic inference. *PLoS computational biology*, 12(3):e1004845.

Sjödin, P., Kaj, I., Krone, S., Lascoux, M., and Nordborg, M. (2005). On the meaning and existence of an effective population size. *Genetics*, 169(2):1061–1070.

Smith, J. M. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetics Research*, 23(1):23–35.

Walczak, A. M., Nicolaisen, L. E., Plotkin, J. B., and Desai, M. M. (2012). The structure of genealogies in the presence of purifying selection: a fitness-class coalescent. *Genetics*, 190(2):753–779.

Walsh, B. and Lynch, M. (2018). *Evolution and selection of quantitative traits*. Oxford University Press.

Zeng, K. and Charlesworth, B. (2011). The joint effects of background selection and genetic recombination on local gene genealogies. *Genetics*, 189(1):251–266.