Estimation de l'histoire démographique des populations à partir de génomes entièrement séquencés

Willy Rodríguez

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

l'Institut National des Sciences Appliquées de Toulouse (INSA de Toulouse)

Présentée et soutenue le *20/06/2016* **par :**

WILLY RODRÍGUEZ VALCARCE

Estimation de l'histoire démographique des populations à partir de génomes entièrement séquencés.

JURY

BÉATRICE LAURENT	Professeur d'Université	Président du Jury
OLIVIER FRANÇOIS	Professeur d'Université	Membre du Jury
MARK BEAUMONT	Professeur d'Université	Membre du Jury
LOUNÈS CHIKHI	Directeur de Recherche	Membre du Jury
OLIVIER MAZET	Maître de Conférences	Membre du Jury
RASMUS HELLER	Chargé de Recherche	Membre du Jury

École doctorale et spécialité :

MITT : Domaine Mathématiques : Mathématiques appliquées

Unité de Recherche :

Institut de Mathématiques de Toulouse (UMR 5219)

Directeur(s) de Thèse :

Olivier MAZET et *Lounès CHIKHI*

Rapporteurs :

Olivier FRANÇOIS et *Mark BEAUMONT*

Abstract

The rapid development of DNA sequencing technologies is expanding the horizons of population genetic studies. It is expected that genomic data will increase our ability to reconstruct the history of populations. While this increase in genetic information will likely help biologists and anthropologists to reconstruct the demographic history of populations, it also poses big challenges. In some cases, simplicity of the model could lead to erroneous conclusions about the population under study. Recent works have shown that DNA patterns expected in individuals coming from structured populations correspond with those of unstructured populations with changes in size through time. As a consequence it is often difficult to determine whether demographic events such as expansions or contractions (bottlenecks) inferred from genetic data are real or due to the fact that populations are structured in nature. Moreover, most methods allowing to reconstruct past population size changes do not always account for structure effects. In this thesis, some recent results in population genetics are presented: (i) a model choice procedure is proposed to distinguish one simple scenario of population size change from one of structured population, based on the distribution of coalescence times of two genes, showing that for these simple cases, it is possible to distinguish both models using genetic information of one single individual; (ii) by using the notion of instantaneous coalescent rate, it is demonstrated that for any scenario incorporating structured population, regardless of the complexity, there always exists a panmitic scenario with a precise function of population size changes having exactly the same distribution for the coalescence times of two genes. This not only explains why spurious signals of bottlenecks can be found in structured populations but also predicts the demographic history that actual inference methods are likely to reconstruct when applied to non panmitic populations. Finally, (iii) a method based on a Markov process is developed for inferring past demographic events taking the structure into account. This method uses the distribution of coalescence times of two genes to detect past demographic changes in structured populations from the DNA of one single individual. Some applications of the model to genomic data are discussed.

Contents

Introduction générale

L'objectif principal de cette thèse est l'étude et le développement de modèles permettant de reconstruire certains aspects du passé d'une population à partir de données génétiques. Un des aspects fondamentaux de notre recherche est la reconstruction de l'histoire démographique. L'histoire démographique d'une population peut être caractérisée par des changements de taille, par l'existence de flux de gènes avec d'autres populations, dans le cadre de modèles de populations structurées ou encore par l'existence de phénomènes d'extinctions et de recolonisations dans le cadre de modèles de « métapopulations » (Hey and Machado, 2003). Au cours de cette introduction nous nous concentrerons principalement sur des modèles non structurées et le terme « histoire démographique » sera principalement et fondamentalement caractérisée par les changements de « taille efficace ». Nous entendons par « taille efficace », le nombre d'individus présents dans une population idéale (modèle de Wright-Fisher, décrit plus bas) dont une certaine mesure de la diversité génétique est la même que celle de la population étudiée (voir Charlesworth (2009) pour une discussion sur les différentes notions de taille efficace). Une population idéale est celle qui vérifie les hypothèses du modèle proposé par Wright en 1931, et que nous présentons dans la section suivante. Différentes méthodes statistiques sont utilisées pour estimer la taille efficace d'une population, ainsi que pour inférer la manière dont cette taille efficace change au cours du temps.

L'émergence de nouvelles techniques de séquençage (en anglais *Next Generation Sequencing*) a entraîné un développement accéléré de la génomique des populations. Avec l'augmentation du volume des données disponibles, l'emploi de modèles robustes, capables de tirer un maximum d'information des séquences d'ADN est devenu davantage nécessaire. De nouvelles techniques et de nouveaux modèles ont vu le jour et leur développement continue encore aujourd'hui. Actuellement, il est possible de récupérer entièrement la séquence d'ADN d'un seul individu diploïde et, à partir ce cela, d'estimer certains paramètres de l'histoire démographique de la population d'où il provient. Cela a un impact retentissant, notamment dû au fait que ce genre de méthodes peut aider à mieux connaître l'évolution récente de l'espèce humaine. De plus, la reconstruction de l'histoire démographique est aussi utilisée pour mieux comprendre l'histoire récente des espèces menacées en relation

avec l'histoire des peuplements humains, de leurs impacts sur les environnements naturels ou avec les changements climatiques passés.

La grande majorité des méthodes utilisées aujourd'hui est basée sur des modèles qui font certaines hypothèses simplificatrices sur les populations étudiées. C'est grâce à ces hypothèses simplificatrices qu'il est possible d'appliquer les résultats issus de la théorie des probabilités, afin de décrire l'évolution au cours du temps des populations étudiées. Il est donc très important que le modèle arrive à identifier les caractéristiques fondamentales de la population que l'on veut décrire. Une hypothèse qui est souvent faite est celle du random-mating (ou panmixie), c'est-à-dire que l'on suppose que la reproduction entre les individus se fait de manière aléatoire, et que tous ont la même chance d'avoir des descendants. Cette hypothèse engendre des conséquences qui seront discutées tout au long de cette thèse. Un autre modèle très étudié et particulièrement simple est celui qui suppose que la population est divisée en différentes colonies, entre lesquelles il existe un flux de gènes symétrique dû, par exemple, à la migration d'une certaine proportion d'individus. Ce modèle est nommé *n-island model* (ou modèle en île).

Quelques études ont montré que les modèles utilisés pour décrire l'évolution au cours du temps d'une population sous l'hypothèse de panmixie, peuvent ne pas être appropriés pour décrire l'évolution d'une population structurée. En fait, des résultats de simulations indiquent que, lorsqu'on utilise une méthode basée sur l'hypothèse de panmixie pour reconstruire l'histoire démographique d'une population structurée, on est amené à inférer des changements de taille n'ayant jamais eu lieu. Par ailleurs, l'histoire démographique inférée peut varier, selon la manière de constituer l'échantillon utilisé pour l'analyse. Cela présente un vrai problème au moment d'appliquer un modèle sur un scénario réel. Si nous ne savons pas si la population étudiée est plus proche d'un modèle panmictique que d'un modèle structuré, comment alors interpréter l'histoire démographique inférée par la méthode ? Est-ce que les changements de taille détectés pourraient aussi être expliqués par des effets de la structure ? Quels sont les effets de la structure lorsqu'on veut reconstruire l'histoire démographique sous l'hypothèse de panmixie ?

Le travail de recherche présenté dans cette thèse vise à donner des arguments théoriques pour répondre à ces questions. Cela nous amènera à construire un modèle permettant de détecter correctement les évènements démographiques dans le passé des populations structurées.

Nous commencerons, dans le chapitre 1, par décrire quelques modèles classiques de génétique des populations, qui servent à établir les bases du développement théorique fait dans cette thèse. Nous présenterons le modèle de Wright-Fisher ainsi que des notions telles que la dérive génétique et la panmixie. Nous verrons aussi comment le coalescent de Kingman est obtenu à partir du modèle de Wright-Fisher, lorsque la taille de la population devient grande. Des concepts importants

comme « Ancêtre Commun le Plus Récent » et « temps de coalescence » sont aussi introduits. En particulier, le « temps de coalescence de deux gènes » est la pierre angulaire de tout notre travail. Dans la section 1.3 nous abordons quelques extensions du coalescent de Kingman qui sont fondamentales pour les méthodes visant à inférer l'histoire démographique. Parmi ces extensions, le coalescent structuré s'avère particulièrement important pour nous, car c'est le point de départ pour construire le modèle présenté dans le chapitre 4.

La plupart des méthodes utilisées pour inférer les changements de taille d'une population au cours du temps se basent sur la théorie développée par Griffiths and Tavaré (1994). Cette théorie établit une manière de reconstruire l'histoire démographique *backward*, c'est-à-dire, en remontant le temps du présent vers le passé. On suppose que la taille de la population au présent ($t = 0$) est égale à N_0 et on définit une fonction λ qui permet de calculer la taille à chaque instant $t > 0$ par :

$$N(t) = N_0\lambda(t).$$

Si on prend deux individus haploides (ou deux gènes) au hasard dans la population et qu'on remonte le temps, on finit par trouver à un moment donné, l'ancêtre commun de ces deux individus. On note T_2 le temps auquel l'ancêtre commun de deux gènes apparaît, lorsqu'on remonte le temps du présent vers le passé. On pourra considérer T_2 comme étant une variable aléatoire à valeurs dans $\mathbb{R}_+$. Selon le modèle proposé par Griffiths and Tavaré (1994), dans une population **panmictique** dont la taille change au cours du temps d'une manière déterministe, donnée par une fonction λ, il est possible d'écrire la loi de T_2 comme :

$$F_{T_2}(t) = \mathbb{P}(T_2 \leq t) = \exp\Big(- \int_0^t \frac{1}{\lambda(u)} du \Big).$$

Il est par ailleurs possible de d'établir (sous certaines hypothèses) des relations entre les valeurs de T_2 et les données issues du séquençage de l'ADN, ce qui permet d'appliquer le modèle théorique sur des populations réelles. Néanmoins, nous nous concentrerons dans cette thèse sur l'étude de la distribution de T_2 (le temps de coalescence de deux gènes qui n'est jamais directement accessible avec des « données réelles »).

Il est important de remarquer que la distribution du temps de coalescence sous un modèle de population avec taille variable proposée par Griffiths and Tavaré (1994), est obtenue sous l'hypothèse que la population est **panmictique**. Par des résultats de simulations, il a été montré que si on applique une méthode d'inférence basée sur l'hypothèse de panmixie sur une population structurée, on trouve des changements de taille de population, même si la population n'a pas changé de taille. Cela met en cause l'existence des changements inférés par ce genre de

méthodes. Par conséquent, il est nécessaire de développer des théories permettant de distinguer les vrais changements de taille des effets de la structure. C'est dans cette direction que le travail présenté dans le chapitre 2 est orienté. Nous montrons qu'il est possible de distinguer deux modèles simples, dont un panmictique avec changement de taille et l'autre considérant une population structurée, à partir de la distribution de T_2. Les deux modèles comparés sont très simples :

- Modèle 1 : population panmictique, avec un changement de taille instantané d'un rapport α, survenu à l'instant T (appelé *SSPSC*)
- Modèle 2 : *n-island model* avec un nombre n d'îles et un flux de gènes égal à M (appelé *StSI*)

Chacun de ces deux modèles est gouverné par deux paramètres : (α, T) pour le premier, et (n, M) pour le second. Nous remarquons qu'il existe des jeux de paramètres qui font que les deux premiers moments de T_2 soient très proches sous les deux modèles. Néanmoins, les fonctions de densité de T_2 pour chaque modèle restent assez différentes. Afin de trouver les jeux de paramètres qui font que les deux modèles soient le plus proches possibles (vis à vis de la distribution de T_2), nous mettons en place une stratégie d'estimation par maximum de vraisemblance, à partir d'un vecteur de valeurs de T_2. Des résultats des simulations montrent que cette stratégie est capable de trouver correctement les paramètres correspondant à chaque modèle. Afin de déterminer lequel des deux modèles correspond le plus à un ensemble de valeurs de T_2, nous appliquons la procédure suivante :

1. Estimer par maximum de vraisemblance les paramètres pour chacun des deux modèles à partir des valeurs de T_2 (on utilisera seulement la moitié des valeurs pour des raison d'indépendance entre estimation de paramètres et test statistique)
2. Pour chaque modèle, réaliser un test de Kolmogorov-Smirnov pour déterminer si les valeurs de T_2 correspondent au modèle ou non (on utilisera l'autre moitié des valeurs).

Si, après avoir appliqué le test, un modèle est rejeté et pas l'autre, cela veut dire que le modèle qui n'a pas été rejeté est celui qui correspond le mieux aux données. Si aucun des deux est rejeté, ou les deux sont rejetés, le modèle le plus approprié est choisi par un critère basé sur la valeur de la vraisemblance des paramètres.

Les résultats de simulations sur plusieurs combinaisons de paramètres, indiquent que la stratégie proposée permet d'identifier correctement le modèle le plus approprié pour expliquer les données, à partir des valeurs de T_2. Nous constatons que les paramètres sont très bien estimés, notamment le nombre d'îles

sous un modèle structuré. Cela suggère qu'il doit être possible d'estimer le nombre d'îles sous un modèle de population structurée, à partir d'un seul individu diploïde. Nous discutons aussi comment cette stratégie de choix de modèle peut être appliquée à de données génomiques. À la fin du chapitre (Subsection 2.7.4), nous montrons quelques résultats préliminaires issus de simulations, qui nous font penser que la méthode peut être appliquée sur des scénarios réels, même si une étude approfondie est encore nécessaire. Nous mentionnons également quelques résultats de robustesse qui ont été réalisés dans le cadre d'un stage de M1 (Alexandre Changenet).

Nous avons aussi trouvé une relation intéressante entre les paramètres des deux modèles. En particulier, le nombre d'îles (n) et le ratio du changement de taille (α) sont très corrélés. Cela veut dire, par exemple, que si on applique le modèle 1 sur des données qui correspondent au modèle 2, le ratio du changement de taille inféré sera plus important si le nombre d'îles du modèle 2 est grand. Donc, même un très fort changement de taille inféré par une méthode qui suppose que la population est panmictique, pourrait être faux si la population étudiée est structurée et composée d'un grand nombre d'îles.

La stratégie développée dans le chapitre 2 vise à donner des outils théoriques, afin de déterminer dans quels cas les changements de taille qui apparaissent lorsqu'on reconstruit l'histoire démographique d'une population, sont simplement une conséquence du fait que la population est structurée, et donc, ne sont pas de vrais changements de taille. Néanmoins, les modèles comparés sont très simples. Actuellement, il existe différentes méthodes capables d'inférer non pas un, mais plusieurs changements de taille survenus dans le passé. Il est aussi possible de considérer que la taille d'une population change suivant une fonction λ (par exemple, de manière linéaire ou exponentielle), ce qui implique que la taille de la population est différente à chaque instant t.

Un premier pas en direction d'une généralisation de la méthode à des scénarios plus complexes serait d'étudier s'il est possible de distinguer un modèle de population structurée, d'un modèle considérant des changements arbitraires dans une population panmictique. C'est la question abordée dans le chapitre 3. Nous montrons que pour n'importe quelles valeurs des paramètres (n_0, M_0) sous un *n-island model*, il est toujours possible de trouver une fonction λ telle que la fonction de répartition de T_2, sous un modèle panmictique, avec une histoire démographique déterminée par λ est *identique* à la fonction de répartition de T_2 sous un *n-island model* avec n_0 îles et un taux de migration égal à M_0. Une conséquence directe de ce résultat est que les valeurs de T_2 ne permettent pas de décider, dans le cas général, si les changements de taille sont dus aux effets de la structure, et donc que le *n-island model* est indiscernable d'un modèle panmictique avec changements de

taille arbitraires.

Par ailleurs, nous montrons comment la fonction λ peut s'exprimer de manière simple en fonction des paramètres du *n-island model*. Cette fonction détermine la manière dont la taille d'une population panmictique doit changer, pour faire en sorte que la distribution de T_2 soit identique à celle du *n-island model* correspondant. Dans un contexte plus général, nous présentons des arguments permettant de voir que quel que soit le modèle considéré, il est possible de décrire un modèle panmictique avec une fonction précise de changements de taille λ dont la fonction de répartition de T_2 est identique à celle du premier. Nous montrons que cette fonction λ peut être obtenue à partir de la fonction de répartition et la densité de T_2 par la relation :

$$\lambda(t) = \frac{\mathbb{P}(T_2 > t)}{f_{T_2}(t)},$$

où f_{T_2} est la densité de T_2. Cette expression s'avère très utile car elle permet de prédire l'histoire démographique reconstruite par n'importe quelle méthode basée sur l'hypothèse de panmixie, lorsque cette méthode est appliquée sur une population qui est proche d'un *n-island model*, et plus généralement d'un modèle quelconque dont on connaît la fonction de répartition et la densité de T_2.

Il est important de remarquer que sous un modèle structuré comme le *n-island model*, cette fonction λ ne correspond pas à des changements de taille. Nous montrerons que sous le *n-island model*, les valeurs de λ peuvent varier même si la taille de la population reste constante, ou peuvent même indiquer une décroissance alors que la taille de la population a augmenté. Par conséquent, les valeurs de λ ne doivent pas être interprétées comme des changements de taille tant qu'on n'a pas vérifié que la population étudiée est proche d'un modèle panmictique. Dans le cadre d'un scénario plus général que celui considérant une population panmictique, l'inverse de la fonction λ (c'est-à-dire, $1/\lambda$) représente le taux de coalescence de deux gènes à chaque instant t. C'est pour cette raison que nous utilisons le terme *IICR* (de l'anglais *Inverse Instantaneous Coalescence Rate*) pour désigner la fonction λ.

Nous pouvons donc constater que la notion de « taille efficace » d'une population sous le *n-island model* devient problématique, car il n'est plus possible de caractériser cette taille par une quantité fixe, même si la taille de la population reste constante. Nous verrons aussi que lorsqu'on remonte suffisamment loin dans le temps sous un *n-island model*, l'*IICR* s'approche d'une asymptote horizontale, qui pourrait être considérée comme une « taille efficace de population ancestrale ». De même, pour de grandes valeurs de M (le taux de migration) dans un *n-island model*, l'*IICR* s'approche rapidement d'une asymptote horizontale qui correspond à ce qui a été considéré par d'autres auteurs comme étant la « taille efficace » d'un modèle structuré, avec un taux de migration très fort (*strong migration limit*).

Une extension simple du *n-island model* consiste à considérer différentes valeurs du taux de migration (M) à différents moments du passé, tout en gardant le nombre d'îles (n) constant. Nous verrons que ces changements de M font varier l'IICR d'une manière particulière : une grande valeur de M dans un intervalle temporel correspond à une valeur plus petite de l'IICR, alors qu'une petite valeur de M fait correspondre un IICR supérieur. Cela implique que lorsqu'on utilise une méthode qui suppose que la population est panmictique alors que la population est en réalité structurée , les changements dans le taux de migration sont interprétés par la méthode comme étant des changements de taille de la population. Ce constat amène à imaginer qu'il est possible de reproduire les changements de taille inférés par une méthode basée sur l'hypothèse de panmixie, en appliquant cette méthode sur une population de taille constante, qui suit un *n-island model*, et dont le taux de migration change. Nous donnons un exemple, considérant une population de taille constante, structurée selon un *n-island model* avec trois valeurs différentes du taux de migration, pendant trois intervalles du passé. Nous montrons que l'histoire démographique reconstruite par une des méthodes les plus récentes (Li and Durbin, 2011) à partir des données simulées sous le scénario décrit, correspond à l'histoire démographique reconstruite par le PSMC à partir des vraies données issues du génome humain.

Les méthodes se basant sur l'hypothèse de panmixie sont les plus utilisées pour reconstruire l'histoire démographique. Cependant, si on les applique sur une population structurée, il existe un risque de retrouver de faux changements de taille qui peuvent être fortement influencés par des changements de taux de migration. Afin de mieux comprendre les effets des changements de taux de migration dans l'histoire démographique reconstruite par ces méthodes (qui n'est autre que l'*IICR*), nous nous intéressons à la fonction de répartition de T_2, sous un *n-island model* avec des changements de taux de migration. Nous proposerons dans le chapitre 4 un modèle basé sur un processus de Markov pour décrire l'évolution de deux lignées sous le *n-island model*. Ce modèle permet de suivre deux lignées, correspondant à deux gènes, en remontant le temps, du présent ($t = 0$) jusqu'au moment où on trouve leur ancêtre commun (T_2). Nous appelons ce modèle NIMC (en anglais *N-Islands Markov Chain*). Le NIMC calcule la fonction de répartition de T_2 sous un modèle de population structurée (le *n-island model*), permettant notamment d'inclure des changements dans le taux de migration à différents instants dans le passé.

Le modèle à partir duquel nous construisons le NIMC est un processus de Markov à temps continu. Ce processus considère trois états différents. À un instant t donné, les deux lignées peuvent :

1. être dans la même île

2. être dans des îles différents

3. avoir atteint leur ancêtre commun.

Nous construisons le générateur infinitésimal de ce processus sous un *n-island model*, ce qui permet de calculer le semigroupe de transition associé. La matrice du semigroupe de transition associé à ce processus contient la fonction de répartition de T_2 ainsi que sa densité.

La propriété de Markov dans le NIMC permet de décrire l'évolution des deux lignées jusqu'à l'apparition de l'ancêtre commun, à partir de n'importe quel instant t, indépendamment de l'évolution du processus avant t. Nous utilisons cette propriété pour introduire un changement dans le taux de migration à l'instant $t = t_1$. Il est de même possible de changer le taux de migration à plusieurs instants différents $(t_1, t_2, ..., t_n)$. Par ailleurs, le même argument permet de rajouter des variations dans la taille totale de la population à différents instants du passé.

Il est donc possible, dans le cadre du modèle NIMC, de calculer la fonction de répartition ainsi que la densité de T_2 sous un *n-island model* avec des changements de taux de migration et de changements de taille.

Le modèle que nous développons dans le chapitre 4 ouvre différentes voies pour de futures études en génétique des populations. La possibilité de connaître la fonction de répartition ainsi que la densité de T_2 permet, par exemple, de prédire l'histoire démographique qui sera reconstruite par une méthode basée sur l'hypothèse de panmixie, lorsqu'elle est appliquée sur une population structurée. Par ailleurs, dans le cadre du NIMC, il devient possible d'étudier l'histoire démographique d'une population structurée, tout en éliminant les effets de faux changements de taille présentes lorsqu'on utilise une méthode supposant que la population est panmictique.

Chapter 1

Models and concepts

The present chapter contains a description of some fundamental models and concepts from population genetics. The contributions presented in this thesis heavily rely on these models and concepts. Some important definitions such as *panmictic population* and *genetic drift* are introduced in the models presented in section 1.1. A key population genetic model is presented in section 1.2: *the coalescent*, along with the concept of *coalescence time*. In particular, the *coalescence time of two genes* is the cornerstone of this thesis. The different extensions of the coalescent presented in section 1.3 are fundamental for many methods used for reconstructing the demographic history of populations. Especially, the *structured coalescent* is very important, given that it provides the basis for the model developed in chapter 4. Readers familiar with these concepts may still want to read this section even if superficially to accustom themselves with the terminology, and formalism.

1.1 Early models on population genetics

1.1.1 Hardy-Weinberg equilibrium

The origins of the mathematical theory of population genetics can be traced back to 1908 with the works of the British mathematician Godfrey Harold Hardy (Hardy, 1908) and the German physician Wilhelm Weinberg (Weinberg, 1908). Their papers where independently published, in English and German respectively, withing a few months of each other. Nowadays, the law stated by them is know as the **Hardy-Weinberg equilibrium**. For some anecdotal and historical comments, see Crow (1988). The derivation of the law can be presented as follows.

Consider a diploid population evolving in the absence of any evolutionary forces (i.e. there is no selection, no mutation and mating between individuals occurs at random without any kind of reproductive advantage). If we look at some partic-

	AA	**Aa**	**aa**
AA	AA	$\frac{1}{2}AA$; $\frac{1}{2}Aa$	Aa
Aa	$\frac{1}{2}AA$; $\frac{1}{2}Aa$	$\frac{1}{4}AA$; $\frac{1}{2}Aa$; $\frac{1}{4}aa$	$\frac{1}{2}Aa$; $\frac{1}{2}aa$
aa	Aa	$\frac{1}{2}Aa$; $\frac{1}{2}aa$	aa

Table 1.1: Frequencies of alleles after random mating

ular locus where any of two alleles (A or a) may be present, we have then three possibilities, AA for dominant (homozygous), Aa for heterozygous and aa for recessive (homozygous). Suppose that at some generation g, the frequencies of AA, Aa, aa are respectively $p, 2q, r$ and that sexes are evenly distributed over the three variants. In addition let's assume that generations are not overlapping so that, when giving birth to a new generation, the current one disappears. Under these assumptions, the allele frequencies $p_1, 2q_1, r_1$ at generation $g+1$ can be calculated according to the frequencies at generation g. Considering all possible pairings (table 1.1.1) we get:

$$\begin{aligned} p_1 &= (p+q)^2 \\ 2q_1 &= 2(p+q)(q+r) \\ r_1 &= (q+r)^2 \end{aligned} \tag{1.1}$$

It is worth commenting that in order to carry out the computations for obtaining the relations in equation 1.1 it has been implicitly assumed that population size (denoted N) and, consequently, the number of individuals having one specific allele, is so big that, for example, fractions N_{AA}/N and $(N_{AA}-1)/(N-1)$ are both equal to p, which is the frequency of AA (here N_{AA} represents the number of individuals having the genotype AA). The question of interest here is whether alleles frequencies in generation $g+1$ are the same as in generation g. By looking at $(2q_1)^2$ and recalling that $p+2q+r=1$, it can then be deduced from equation 1.1 that the condition for proportions to remain unchanged is:

$$q^2 = pr \tag{1.2}$$

as stated in (Hardy, 1908) and (Weinberg, 1908).

In other words, the Hardy-Weinberg law stands that, under the above hypothesis and provided that condition 1.2 holds, alleles frequencies will not change from one generation to the next. Furthermore, this *equilibrium* is maintained, as long as conditions do not change. The reason for this is that $q_1^2 = p_1 r_1$, which implies that alleles frequencies on generation $g+2$ will be equal to those of generation $g+1$

and so on. Note also that $q_1^2 = p_1 r_1$ always holds, no matter what the values of p, q and r are in the previous generation. Consequently, if any external factor (for instance, changes in the environment leading to non random matting or selective pressure) forces the frequency of alleles in the population to move away from the equilibrium, once the influence of external factors disappears, the equilibrium is reached at the next generation, regardless of the proportions of alleles at the time when all conditions were restored. This almost instantaneous move to the equilibrium from any allelic proportions is rather stunning. It should also be noted that there are infinitely many trios $p, 2q, r$ for which Hardy-Weinberg equilibrium can be reached (see Figure 1.1).

This rather simple principle had a high impact in the biologist community because, even if it describes an idealised state, it overthrew the erroneous idea that the frequency of one allele in the population was determined by whether the allele is dominant or recessive (Hardy, 1908). For a generalisation of the Hardy-Weinberg law to more than two alleles and more than one locus, see Ewens (2012).

1.1.2 The Wright-Fisher model

The Wright-Fisher model, named after the works of Wright (Wright, 1931) and Fisher (Fisher, 1930), is a widely studied model in population genetics. Essentially, the hypothesis are the same as those of Hardy-Weinberg equilibrium with the difference being that the population is finite. This apparently minor difference has important consequences, especially the possibility that one allele may completely disappear from the population without selection. As in the above paragraph, it is assumed that there is no mutation, individuals have equal chances to reproduce, generations are not overlapping, so at every new generation, the entire population is fully replaced by its descendants. In contrast, population size is **finite** (let's say equal to N haploid individuals) and **constant** from one generation to the next.

Under this model, given that all individuals have the same chance to reproduce, the ancestors of individuals in the present generation can be obtained by randomly sampling, with replacement, the individuals of the previous generation (Figure 1.2 left). If we now focus on a single loci with two allelic variants (A and a), it is possible to describe how the proportions of these alleles vary in the population when going forward in time. Note that, at generation $n + 1$, the probability that one individual has allele A is equal to the proportion of allele A at generation n. Denoting p the proportion of allele A in generation n, any individual at generation $n + 1$ has allele A with probability p. Consequently, the number of individuals having the allele A at generation $n+1$ can be modelled by a Binomial distribution with parameters N and p, where N is the size of the population. Note that the expected value for the number of individuals with allele A in generation $n+1$ (the expected value of a Binomial distribution is Np) is equal to the number of individ-

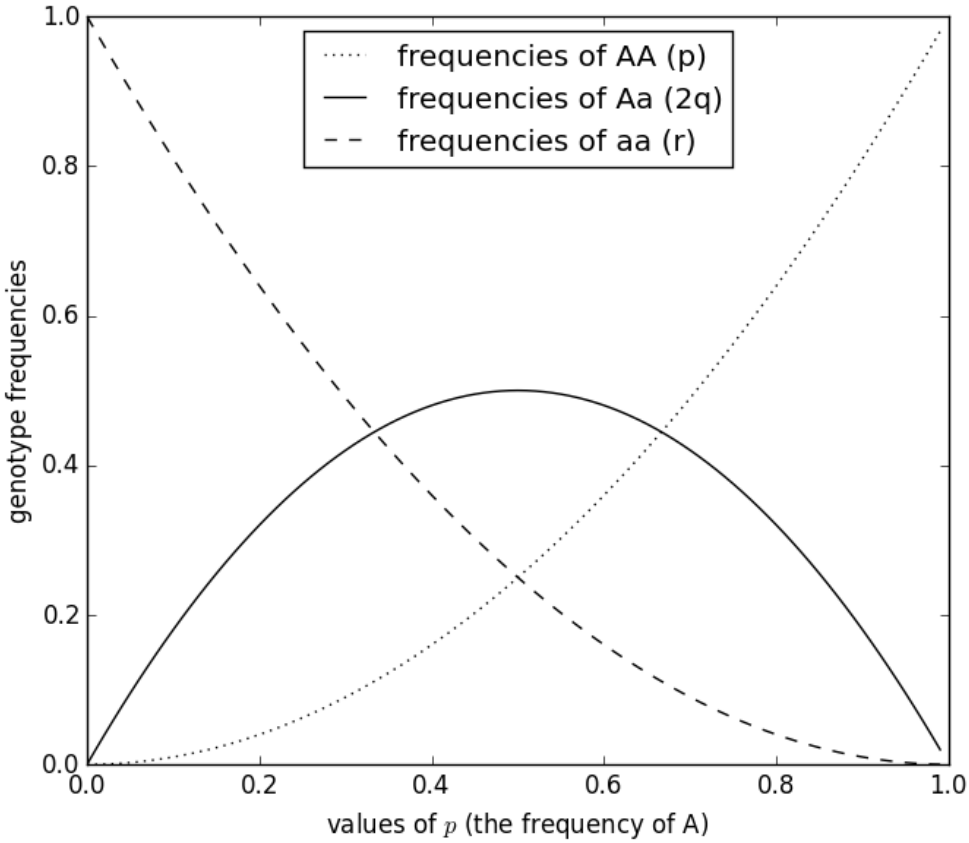


Figure 1.1: Allele frequencies in Hardy-Weinberg equilibrium. The intersections of any vertical line with these three curves indicates values of $p, 2q$ and r satisfying the condition of Hardy-Weinberg equilibrium.

uals with allele A in generation n. This is a similarity with the Hardy-Weinberg equilibrium: on average, allele frequencies are expected to remain constant. However, the fact that population size is finite makes the number of individuals with allele A to vary randomly following a Binomial distribution. In other words, in a finite population allele frequencies are subject to *genetic drift* (Figure 1.3 a). Note also that at each generation there is a non zero probability that one of the two alleles becomes **extinct**, which means that the other one has become **fixed** in the population. Another interesting point is that the quantity of individuals with allele A at the next generation depends only on the proportion of allele A in the current generation, with no dependency on allelic proportions before. This allows to see the number of individuals in the population having the allele A at each

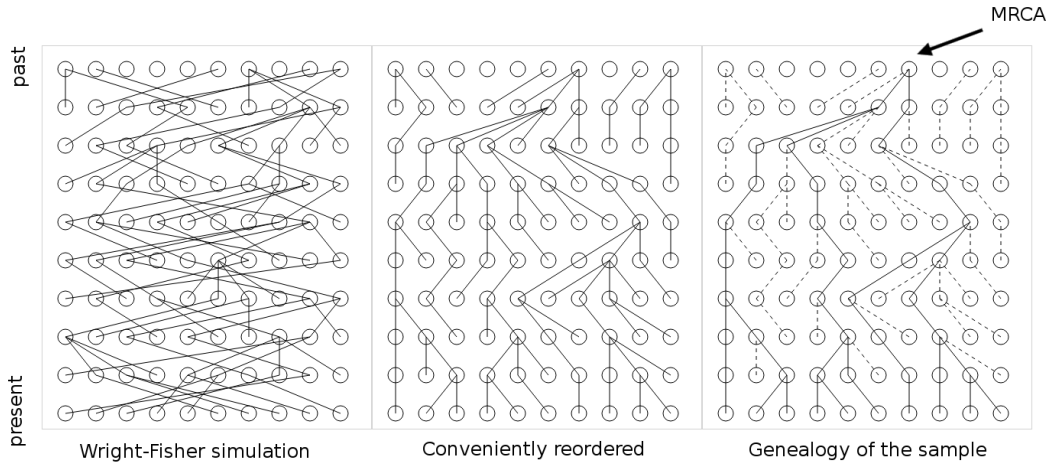


Figure 1.2: Wright-Fisher simulation and genealogy of the sample at the present

generation, as a Markov Chain with states space $E = \{0, 1, ..., N\}$ and transition probability given by:

$$P_{kl} = \mathbb{P}(N_A^{i+1} = k | N_A^i = l) = \binom{N}{k} \Big(\frac{l}{N}\Big)^k \Big(1 - \frac{l}{N}\Big)^{(N-k)} \tag{1.3}$$

for any k and l elements of E and for N_A^i being the number of individuals with allele A at generation i. Note also that if the number of individuals with allele A reaches the value zero, it will remain zero thereafter. The same occurs if it gets the value N. Using the Markov chain terminology, this means that states 0 and N are *absorbing states*.

Another interesting point is that under the Wright-Fisher model, diversity in the population disappears almost certainly. It can be seen from equation 1.3 that the probability of having only individuals with allele A at any generation is not zero and so is the probability of not having any individual with allele A. Denoting τ the number of the generation (going forward in time) when allele A either gets fixed or disappears, it can be shown that the probability of τ to be finite is equal to one. For technical details see Delmas and Jourdain (2006) and Ewens (2012). This means that one of the two alleles will be fixed in the population with probability one and that diversity will disappear. It can also be proved that the probability that allele A gets fixed in the population is equal to the proportion of allele A in the first generation. Moreover, the expected time for diversity to disappear tends to a continuous function of the initial proportion of allele A (denoted X_0) as N increases (equation 1.4).

$$\mathbb{E}\Big[\tau | X_0 = \lfloor Nx \rfloor\Big] \sim -2N\Big(xlog(x) + (1-x)log(1-x)\Big) \tag{1.4}$$

In Figure 1.3 (b), the mean values of τ has been computed from independent simulations of generations until fixation or extinction of one allele. For each value of the number of individuals with allele A at the first generation (i.e. $0 < X_0 < 100$), 1000 independent simulations have been done, then we computed mean of observed τ. For the population size we used $N = 100$.

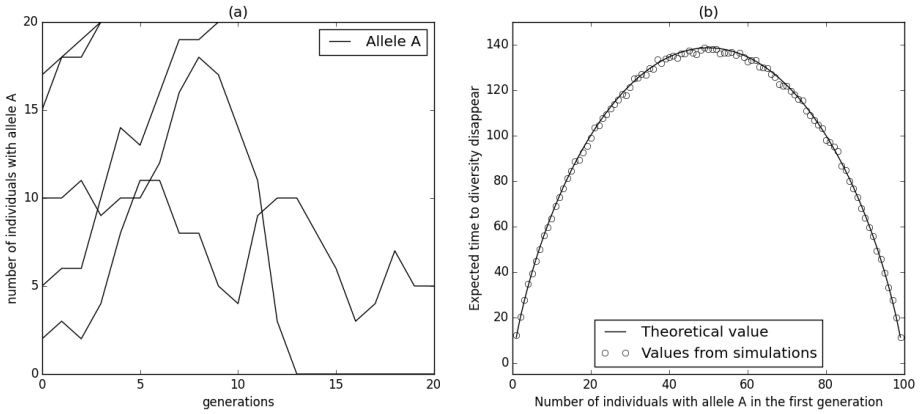


Figure 1.3: (a) Five independent simulations of the number of individuals with allele A in a population of size 20 under the Wright-Fisher model. (b) Empirical and theoretical values of the mean of the time for diversity to disappear in a Wright-Fisher model with population size $N = 100$.

There are other models to study the evolution of populations forward in time. For example, Moran (1958) considers a haploid population in which generations are overlapping. In the Moran model, at times $t = 1, 2, ...$, we choose two individuals randomly and with replacement from the population. One of them reproduces (i.e. copies itself) and the other dies. Each individual has probability $1/N$ of being chosen, with the special case that (given that choice is made with replacement) the same individual can reproduce and then die. Thus, the population size does not change. Unlike in the Wright-Fisher model, the number of individuals with allele A can only increase by one, decrease by one or remain constant after one iteration.

The Moran model and the Wright-Fisher model belong to a class of population genetic models allowing to describe gene genealogies over the time. Under these models, if we trace lineages back in time (Figure 1.2 right) and let the population size to increase (formally $N \to +\infty$), we obtain a random process which is very important in population genetics: **the coalescent**. See Wakeley (2009) for a very

clear derivation of the coalescent from the Wright-Fisher and the Moran models. In the next section we will see a brief description of the coalescent as a limit random process for the Wright-Fisher model when population size increases.

1.2 The coalescent

The coalescent, also known as the Kingman's coalescent after the works of Kingman (Kingman, 1982a,c,b), is a milestone in population genetics. A clear understanding of the principal insights of the coalescent is crucial to perceive the main ideas behind most of the models used nowadays. Basically, the coalescent is a random process allowing to reconstruct the genealogy of a group of haploid individuals sampled in the present (Figure 1.2 right). This genealogy can always be modeled regardless of what happens with the genealogy of the rest of individuals in the population (for example, take just the first four individuals in Figure 1.2 right). When analyzing the genealogy backward in time under a Wright-Fisher model, the assumption that all individuals have the same chance to reproduce is equivalent to consider that each individual "chooses" its ancestor randomly within the previous generation. If, by chance, two individuals choose the same ancestor, we say that a **coalescent event** has occurred. If we start with a sample size of k, then at most $k-1$ coalescent events can occur (note that more than two individuals could choose the same ancestor in the previous generation, which corresponds to a coalescent event between more than two individuals). Note also that after a coalescent event occurs, the number of distinct lineages decreases at least by one. When the number of distinct lineages gets equal to one, we say that the Most Recent Common Ancestor (MRCA) has been reached.

Let's suppose that we sample k individuals from a population evolving under a Wright-Fisher model. We want to reconstruct their genealogical tree. When moving back from one generation to the previous, it is convenient to distinguish three cases: there are k different ancestors in the previous generation ($G_{k,k}$), there are $k-1$ different ancestors (denoted $G_{k,k-1}$), and there are $k-2$ or less different ancestors ($G_{k,l}$, with $l \leq k-2$). In the first case, no coalescent event occurred, in the second, only one coalescent event occurred and in the third there was two or more coalescent events. The probabilities of the first two cases can be computed as follows:

$$\begin{aligned}\mathbb{P}(G_{k,k}) =& \frac{N(N-1)(N-2)...(N-k+1)}{N^k} \\ =& (1-\frac{1}{N})(1-\frac{2}{N})...(1-\frac{k-1}{N}) = 1 - \frac{\sum_{i=1}^{k-1} i}{N} + \mathcal{O}(\frac{1}{N^2}) \\ \mathbb{P}(G_{k,k-1}) =& \frac{\binom{k}{2} N(N-1)(N-2)...(N-k+2)}{N^k} \\ =& \frac{\binom{k}{2}}{N}(1-\frac{1}{N})(1-\frac{2}{N})...(1-\frac{k-2}{N}) = \frac{\binom{k}{2}}{N} + \mathcal{O}(\frac{1}{N^2})\end{aligned} \tag{1.5}$$

where $\mathcal{O}(\frac{1}{N^2})$ is a term that decreases to zero as fast as $\frac{1}{N^2}$ when N goes to infinity. As the identity

$$\mathbb{P}(G_{k,k}) + \mathbb{P}(G_{k,k-1}) + \mathbb{P}(G_{k,l}) = 1$$

must hold, we deduce that $G_{k,l} = \mathcal{O}(\frac{1}{N^2})$. Thus, for high values of N, more precisely when N is much bigger than k ($k << N$), the value of $G_{k,l}$ can be neglected and thus, only two cases are possible (i.e. there is at most one coalescent event when going from one generation to the previous). Of course, this approximation is not accurate for low values of N. For example, if $N = 20$ and $k = 10$ we have that $P(G_{10,8}) = 0.372$ which is too large to be neglected. In Figure 1.4 we can observe how the assumption that no more than one coalescence event occurs in the previous generation, gets more accurate as N increases.

Under this approximation, the process of moving back, one generation at a time, can be simulated by a series of independent random variables with Bernoulli distribution. The probability of success for the Bernoulli is the probability of $G_{k,k-1}$ (the number of different lineages decreases by one). Note also that each success will change the parameter of the Bernoulli thenceforth, while it will remain the same if there is no coalescent event. Thus, for a sample of k individuals, the probability of a coalescent event to occurs when moving one generation back (success) is equal to $k(k-1)/2N$. Consequently, the number of generations we have to move back until the first coalescent event appears (denoted T_k^g) follows a Geometrical distribution with parameter $k(k-1)/2N$. The probability that the k lineages stay distinct for more than τ generations is then computed by:

$$\mathbb{P}(T_k^g > \tau) = \left(1 - \frac{k(k-1)}{2N}\right)^\tau \tag{1.6}$$

Define now a real valued random variable T_k such that $\lfloor NT_k \rfloor = T_k^g$ ($\lfloor x \rfloor$ denotes the integer part of x). In other words T_k is a way of counting the time in units of N generations (when $T_k = 1$, $T_k^g = N$). For any $t \in \mathbb{R}$ we have from equation 1.6:

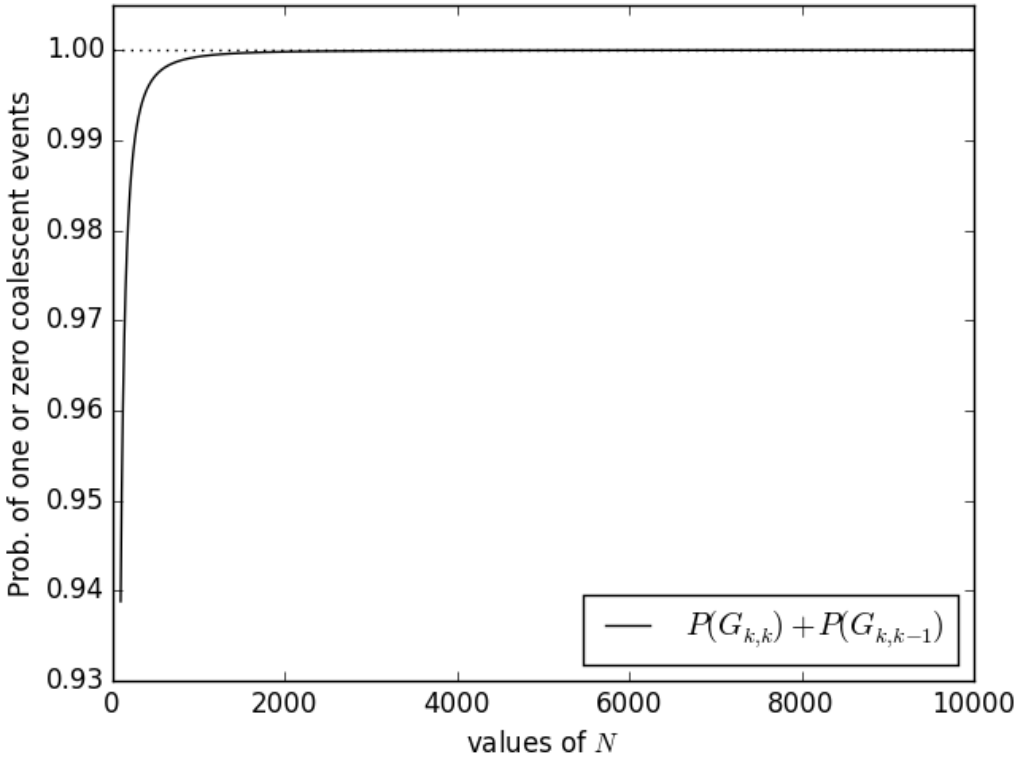


Figure 1.4: Probability of having no more than one coalescent event when moving back to the previous generation under the Wright-Fisher model. The probabilities are computed for a fixed sample size ($k = 10$) and for values of N going from 100 to 10000.

$$\mathbb{P}(T_k > t) = \mathbb{P}(T_k^g > \lfloor Nt \rfloor) = \left(1 - \frac{k(k-1)}{2N}\right)^{\lfloor Nt \rfloor} \tag{1.7}$$

As we are interested in large values of N, using that:

$$\left(1 - \frac{k(k-1)}{2N}\right)^{\lfloor Nt \rfloor} \xrightarrow[N \to +\infty]{} e^{-\frac{k(k-1)}{2}t}$$

the random variable T_k can be considered as an Exponential with parameter $k(k-1)/2$. We have then:

$$F_{T_k}(t) = \mathbb{P}(T_k \leq t) = 1 - e^{-\binom{k}{2}t} \tag{1.8}$$

It is also interesting to comment that, if we focus only on two lineages from the sample of size k, the probability that these two lineages coalesce in the previous

generation is equal to $1/N$. Using the same reasoning as before, the number of generation during which these two lineages will remain different, follows a Geometrical distribution with parameter $1/N$. Then, counting the time in units of N generations, we have that the time during which two particular lineages remain separate (that will be denoted T_2) follows an Exponential distribution of parameter one. This is another way to understand T_k (the time during which k lineages will remain distinct):

$$T_k = \min\{T_2^1, T_2^2, ..., T_2^k\}, \text{ with } T_2^i \sim Exp(1) \text{ for } i = 1...k$$

The minimum of k exponential random variables has an exponential distribution with parameter being the sum of all the parameters of the exponential random variables. We have again that $T_k \sim Exp\Big(k(k-1)/2\Big)$.

By using the coalescent approximation, the genealogy of a sample of size k can be simulated by $k-1$ independent Exponential random variables $\{T_k, T_{k-1}, ..., T_2\}$, with $T_i \sim Exp\Big(i(i-1)/2\Big)$. Each value of T_i represents the time during which i lineages stay different (in units of N generations), that is the time when the first coalescence of i individuals occurs. The lineages that coalesce at each time are chosen randomly from the i different lineages (see Figure 1.5 for an example of simulation).

In a more theoretical way, which is actually the way it was presented by Kingman (1982c), the coalescent is a continuous-time Markov process over a discrete state space. Starting with a sample of size k, the discrete state space (denoted φ) will be the set of all different partitions of $\{1, 2, ..., k\}$. At the present (time zero), we consider the partition formed by all the singletons $\{i\}$, with $i = 1...k$. After the first coalescence event, say it was between i and j, a new partition will be created by joining the sets i and j while keeping the others unchanged. Consider ξ and η, two elements of φ, we denote "$\xi \prec \eta$" if η can be obtained from ξ by joining together two elements of ξ. Consequently, $|\eta| = |\xi| - 1$ ($|\cdot|$ means the number of elements). At any time, the distinct lineages of the genealogical tree correspond to one partition of $\{1, 2, ..., k\}$. After a coalescence event, two lineages are joined as well as the corresponding elements of the partition (see Figure 1.5 right side). The time at which coalescence events occur between two different lineages is exponential with rate one, and whenever two lineages coalesce, the process jumps from one state to another. Thus, the coalescent is a Markov process with Q-matrix (or infinitesimal generator) given by:

$$Q = (q_{\xi\eta}) : q_{\xi\eta} = \begin{cases} -i(i-1)/2 & \text{if } \xi = \eta, \text{ with } i = |\xi| \\ 1 & \text{if } \xi \prec \eta, \\ 0 & \text{otherwise} \end{cases}$$

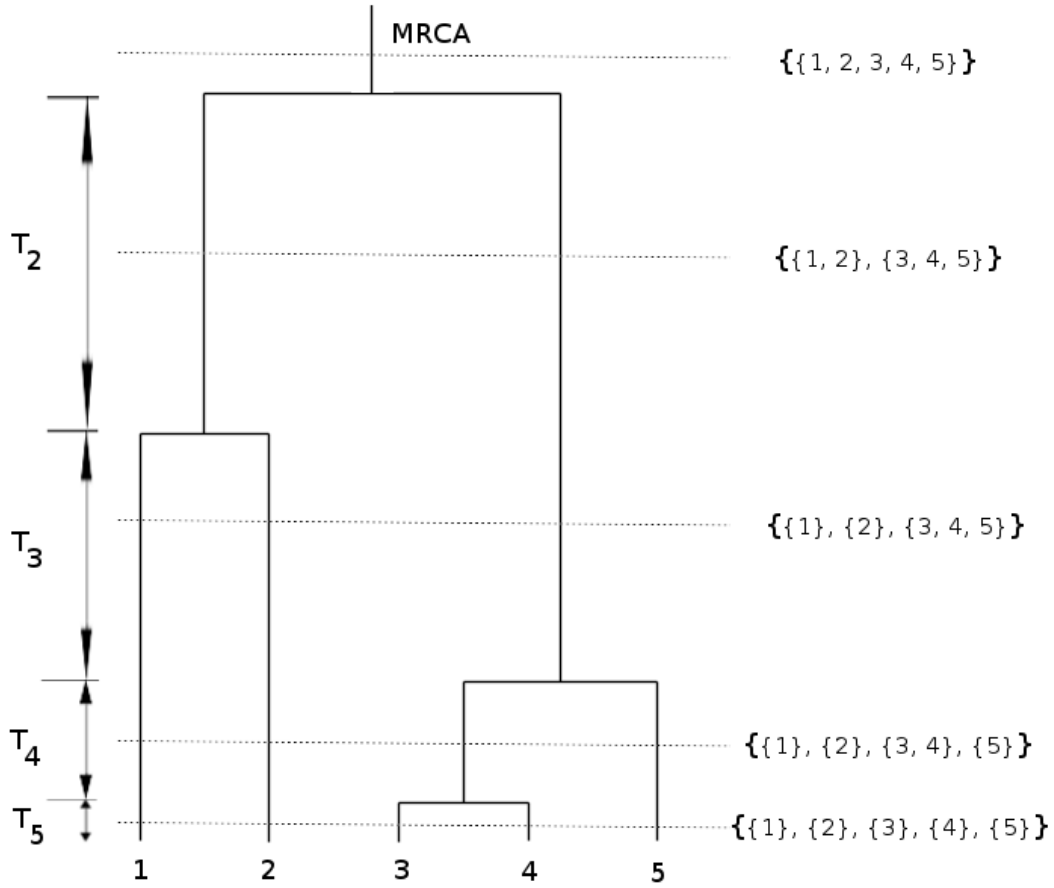


Figure 1.5: Example of a coalescent tree. At the left, the values of coalescence time for $5, 4, 3$ and 2 lineages respectively. At right, we can see how a new partition of the set $\{1, 2, 3, 4, 5\}$ is obtained after a coalescence event by join two subsets.

Example Take only three genes, labeled $\{1\}$, $\{2\}$, and $\{3\}$. The coalescence tree is described by all possible configurations of the three lineages (at time t, $\Big\{\{1\}, \{2, 3\}\Big\}$ means that lineage 2 and 3 have already merged or coalesced while lineage 1 remains distinct). The state space is then given by:

$$\varphi = \Big\{\Big\{\{1\}, \{2\}, \{3\}\Big\}, \Big\{\{1, 2\}, \{3\}\Big\}, \Big\{\{1, 3\}, \{2\}\Big\}, \Big\{\{2, 3\}, \{1\}\Big\}, \Big\{\{1, 2, 3\}\Big\}\Big\}$$

Keeping the states in the same order as above, the Q-matrix is:

$$Q = \begin{pmatrix} -3 & 1 & 1 & 1 & 0 \\ 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

According to the theory of continuous-time Markov process, the corresponding transition semigroup (denoted P_t) is obtained by doing the matrix exponential ($P_t = e^{tQ}$). This is a matrix whose entries ($P_t(i,j)$) represent the probability of being in state j at time t given that the process was in state i at time zero. In this case the transition semigroup P_t is a 5×5 matrix given by:

$$P_t = \begin{pmatrix} e^{-3t} & \frac{e^{-t}-e^{-3t}}{2} & \frac{e^{-t}-e^{-3t}}{2} & \frac{e^{-t}-e^{-3t}}{2} & 1 - \frac{3e^{-t}-e^{-3t}}{2} \\ 0 & e^{-t} & 0 & 0 & 1-e^{-t} \\ 0 & 0 & e^{-t} & 0 & 1-e^{-t} \\ 0 & 0 & 0 & e^{-t} & 1-e^{-t} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The distribution function of the time of the first coalescence event (which is in this case T_3, given that the sample size is equal to three) can be computed using the matrix P_t. The probability that a coalescence event occurs in the interval $[0,t]$ is equal to the probability that the to be in state "two", "three","four" or "five", given that it was in state "one" at time zero. More simply, this probability is equal to the probability of not being in state "one" at time t. We have:

$$\begin{aligned} F_{T_3}(t) = \mathbb{P}(T_3 \leq t) =& P_t(1,2) + P_t(1,3) + P_t(1,4) + P_t(1,5) \\ =& 1 - P_t(1,1) = 1 - e^{-3t} \end{aligned}$$

in correspondence with 1.8.

In a coalescence tree, the length of the branches corresponds to the time intervals between coalescence events. Thus, the total height of the tree is the time when the sample reaches its MRCA. A natural question then is how far in the past should we go in order to find the ancestor of k individuals sampled in the present. Denoting $H_T(k)$ the height of the coalescence tree of k genes (which is actually the time to reach the MRCA of the sample), we can see that $H_T(k)$ is the sum of $k-1$ independent exponential random variables. Hence, it is possible to compute the mean height of the tree by:

$$\mathbb{E}\Big(H_T(k)\Big) = \mathbb{E}\Big(\sum_{i=2}^{k} T_i\Big) = \sum_{i=2}^{k} \mathbb{E}(T_i) = \sum_{i=2}^{k} \frac{2}{i(i-1)} = 2(1 - \frac{1}{k}) \qquad (1.9)$$

From the above formula we can draw two interesting conclusions. First, under a Wright-Fisher model with large population size, the MRCA of a sample, on average, will never be beyond $2N$ generations (N being the population size), regardless of the sample size. This implies that adding individuals to the sample is not expected to increase the height of the corresponding coalescence tree. Second, due to the fact that $\mathbb{E}(T_2) = 1$, the last two lineages are in average those who will take more time to coalesce. In other words, more than half of the coalescence tree will have only two branches on average. This is somehow intuitively because lineages coalesce at rate "k choose 2" which implies that coalescence events are more likely to arrive when there are many lineages, and less likely when there are just a few. Consequently, the variance of the total height of the tree depends more on the variance of T_2 than on the variance of any T_k for $k > 2$. Another interesting result is that the probability that the MRCA of the whole population is reached with a sample of size k is $(k-1)/(k+1)$. See Ian W. Saunders (1984) for more details on this and related results.

1.3 Some extensions of the coalescent

1.3.1 The coalescent with variable population size

Consider that the population size changes over the time in a deterministic way. Denote $N(j)$ the population size j generations before the present. The size of the population at the present is $N(0)$ and it will be used as the reference population size (i.e. $N = N(0)$). Define, for any value of N the function λ_N as:

$$\lambda_N(t) = \frac{N(\lfloor Nt \rfloor + 1)}{N}.$$

This function will be the *population size change function* when using a time scale in units of N generations. We also suppose that the changes in population size are of the same magnitude for any value of N. Formally, we assume that

$$\lim_{N\to\infty} \lambda_N(t) = \lambda(t),$$

where $\lambda(t)$ is *finite* and *strictly positive* for all $t \geq 0$.

Using the same notation as above, T_2^g represents the number of generations we have to move back until the occurrence of a coalescent event between two lineages. Assume population is evolving under a Wright-Fisher model, except that now, the population size may vary at each generation. The probability that two lineages remain different for more than s generations backward in time can be calculated by:

$$\mathbb{P}(T_2^g > s) = \prod_{j=1}^{s} \Big(1 - \frac{1}{N(j)}\Big), \tag{1.10}$$

which gives, when taking the log:

$$-\log\Big(\mathbb{P}(T_2^g > s)\Big) = -\sum_{j=1}^{s} \log\Big(1 - \frac{1}{N(j)}\Big). \tag{1.11}$$

Applying now the inequality

$$\forall x \in [0,1], \quad x \leq -\log(1-x) \leq \frac{x}{1-x},$$

for $x = 1/N(j)$ and summing over all values of j, we get:

$$\sum_{j=1}^{s} \frac{1}{N(j)} \leq -\sum_{j=1}^{s} \log\Big(1 - \frac{1}{N(j)}\Big) \leq \sum_{j=1}^{s} \frac{1}{N(j)-1}. \tag{1.12}$$

Moreover, note that:

$$\sum_{j=1}^{s} \frac{1}{N(j)} = \int_0^s \frac{1}{N(\lfloor u \rfloor + 1)} du$$

and, doing the change of variable $v = u/N$, this integral becomes:

$$= \int_0^{s/N} \frac{N}{N(\lfloor Nv \rfloor + 1)} dv = \int_0^{s/N} \frac{1}{\lambda_N(v)} dv.$$

By doing a similar transformation to the right term of the inequality 1.12, we have for all $s \in \mathbb{R}$ and all value of N:

$$\int_0^{s/N} \frac{1}{\lambda_N(v)} dv \leq -\log\Big(\mathbb{P}(T_2^g > s)\Big) \leq \int_0^{s/N} \frac{1}{\lambda_N(v) - \frac{1}{N}} dv. \tag{1.13}$$

Taking $s = \lfloor Nt \rfloor$ and recalling that T_2 is such that $\lfloor NT_2 \rfloor = T_2^g$, if we take the limit when N goes to infinity we obtain:

$$\int_0^t \frac{1}{\lambda(v)} dv \leq -\log\Big(\mathbb{P}(T_2 > t)\Big) \leq \int_0^t \frac{1}{\lambda(v)} dv$$

and finally:

$$\mathbb{P}(T_2 > t) = \exp\Big(-\int_0^t \frac{1}{\lambda(v)} dv\Big). \tag{1.14}$$

In a similar way, it is possible to compute the distribution of T_k, the probability that k lineages remain distinct for more than t units of coalescent time, under a model with variable population size. The distribution of T_k is given by:

$$\mathbb{P}(T_k > t) = \exp\Big(- \binom{k}{2} \int_0^t \frac{1}{\lambda(v)} dv \Big). \tag{1.15}$$

For more details and some related derivations, see Tavaré (2004).

A population with variable size can also be modeled by a classic coalescent process, but considering a *non-linear* time scale. See Nordborg (2001) for some intuitive explanations. For a rigorous analysis see Donnelly and Tavaré (1995) and Griffiths and Tavaré (1994).

Example The above results are useful for studying many different demographic scenarios. For example, consider a population whose size changes geometrically with rate α. We assume $\alpha \in [-l, l], l > 0$ and $l << N$. Note that, in this example, $\alpha > 0$ means that the population size was lower in the past, while $\alpha < 0$ means that the population size was higher in the past. At each generation, we have:

$$N(j) = \lfloor N(1 - \frac{\alpha}{N})^j \rfloor.$$

The function of population size change is then:

$$\lambda_N(t) = \frac{\lfloor N(1 - \frac{\alpha}{N})^{\lfloor Nt \rfloor + 1} \rfloor}{N}$$

and satisfies that:

$$\frac{N(1 - \frac{\alpha}{N})^{\lfloor Nt \rfloor + 1} - 1}{N} \leq \lambda_N(t) \leq \frac{N(1 - \frac{\alpha}{N})^{\lfloor Nt \rfloor + 1}}{N}$$

when taking the limit we get:

$$\lambda(t) = e^{-\alpha t}.$$

Now, we can compute the distribution of the coalescence time for two lineages using equation 1.14:

$$\mathbb{P}(T_2 > t) = \exp\Big(- \int_0^t e^{\alpha v} dv \Big) = \exp\Big(\frac{1 - e^{\alpha t}}{\alpha} \Big)$$

and by 1.15 we also have:

$$\mathbb{P}(T_k > t) = \exp\Big(\binom{k}{2} \frac{1 - e^{\alpha t}}{\alpha} \Big).$$

From the two above equations we can note that for high values of α, which correspond with a stronger population expansion, it is more likely that lineages remain different for more time. In other words, a larger population size makes the occurrence of coalescent events more difficult. When the present population size is larger than the past population size, coalescent trees are likely to have longer branches at the bottom and many coalescent events will appear at the top (Figure 1.6 right side). Inversely, if the population is small at the present and large in the past, coalescence events between lineages are more likely to occur close to the present, which cause the trees to have shorter branches (Figure 1.6 left side).

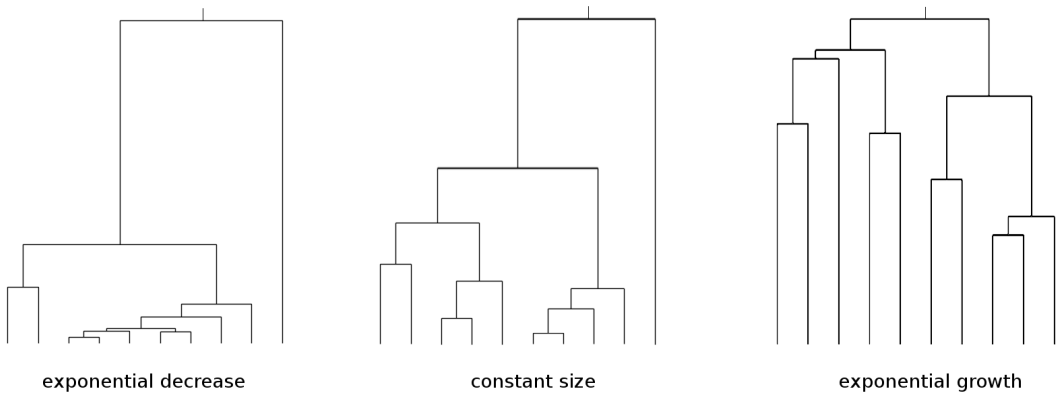


Figure 1.6: Influence of the demographic history on the topology of coalescence trees. Left: population decrease exponentially; center: constant size population; right: exponential growth.

We can think at the function $1/\lambda(v)$ as an *intensity function* or an *instantaneous coalescence rate* because it determines if coalescence events are more likely to occur as t varies. This will be discussed with more details in chapter 3. On the other hand, $\lambda(v)$ corresponds to the function of population size changes in a random-mating population. Most of the methods proposed for reconstructing past demography of populations are based on the development presented above. Under the assumption that the function λ is estimated correctly, it is possible to know the population size at any time in the past by the equality $N(t) = N\lambda(t)$ with N being the present population size which is used as reference. The time can be also be re-scaled in different ways (years, generations, number of substitutions, etc.) but we will not go into the details here.

1.3.2 The structured coalescent

Kingman's coalescent is based on the assumption that populations are panmictic and isolated. However, most populations in nature are subdivided into colonies of different sizes exchanging genetic material through migration of individuals or genes. Models allowing to incorporate this structure are then needed in order to describe more realistic scenarios. However, including structure in the coalescent implies an increasing of complexity. When modelling the genealogy in a subdivided population, it is not enough to trace back only the ancestry of the sample. If we want to maintain the Markovian character of the coalescent, we also need to keep track of the *locations* of lineages at each time in the past. The reason is that, just after the reproduction, the descendants of one single individual are in the same island. Going back in time, this implies that coalescence events between two individuals can only occur if both are in the same island.

A coalescent-like process was introduced by Takahata (1988) for a population subdivided into two colonies. The general form of this ancestral process was subsequently formulated by Notohara (1990). We will see a brief presentation of this process, which be refer to as the "structured coalescent". This presentation can also be found in Herbots (1994) and Wilkinson-Herbots (1998).

Consider a haploid population, divided into colonies or subpopulations which are partially connected between each other by migration. Reproduction inside each colony happens randomly as in a Wright-Fisher model. We identify each subpopulation with subsequent natural numbers, starting from one. Denote $\mathscr{L}$ the set of the subpopulation labels. The size of subpopulation i is $N_i = 2c_iN$ haploid individuals, where c_i is a constant positive integer and N is considered large. The factor two is useful for applications to diploid populations in the sense that, if we assume subpopulation i have c_iN diploid individuals, the number of genes at a particular locus is then $2c_iN$. We sample a finite number n_0 of individuals at some generation (which will be considered time zero), keeping also the information about the island each individual comes from. Then, we count the number of ancestors the sample of n_0 individuals has in each subpopulation, at each time in the past, going from time zero to the time when the MRCA is reached.

As in the standard Kingman's coalescent, the genealogy is then modelled by tracing lineages back in time, but now we also keep track of the island where each lineage is at any time. Coalescence between lineages in different islands are not allowed. When moving back in time, two types of events can occur. The first one is a coalescence between two lineages inside the same island. As a consequence, the number of distinct ancestors the sample has inside that subpopulation decreases by one. It can be proved (Cannings, 1974) that the rate at which this event occurs is inversely proportional to the size of the subpopulation. The other event that can occur when moving back in time can be seen as a "backward migration", which is

when an ancestor of subpopulation i is an immigrant from subpopulation j. In this case, the number of distinct ancestors in subpopulation i decreases by one, while that in subpopulation j increases by one. The rate at which this type of event occurs is given by a migration matrix, whose entries contains the migration rate from subpopulation j to subpopulation i (denoted $M_{ij}/2$), after the corresponding change in time scale (Herbots, 1994). By analogous arguments as in the Kingman's coalescence, the probability of having two coalescence events, two migrations or a coalescence and a migration at the same time is assumed to be zero.

In order to write the Q-matrix of the structured coalescent, some notations need to be introduced:

- $\mathscr{L}$: the set of the subpopulation labels;
- $\alpha_i(t), t \geq 0$: the number of ancestors in the subpopulation i at time t. The time has been appropriately re-scaled (Herbots, 1994) and is counted from the present ($t = 0$) to the MRCA;
- $\alpha(t) = (\alpha_i(t))_{i \in \mathscr{L}}$;
- $(\varepsilon)^i$: the element of $\mathbb{N}^{\mathscr{L}}$ having all components equal to zero except the i-th component which is equal to one:

$$(\varepsilon^i)_j = \delta_{ij} = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{otherwise ;} \end{cases}$$

- if $\alpha(t) = \alpha$, when two lineages in subpopulation i coalesce, the value of $\alpha(t)$ changes to $\alpha - \varepsilon^i$;
- if $\alpha(t) = \alpha$, a "backward migration" from subpopulation i to subpopulation j changes the value of $\alpha(t)$ to $\alpha - \varepsilon^i + \varepsilon^j$.

The structured coalescent is then a continuous-time Markov chain $\{\alpha(t), t \geq 0\}$. The entries of its Q-matrix are given by:

$$Q_{\alpha,\beta} = \begin{cases} \binom{\alpha_i}{2} \frac{1}{c_i} & \text{if } \beta = \alpha - \varepsilon^i \\ \alpha_i \frac{M_{ij}}{2} & \text{if } \beta = \alpha - \varepsilon^i + \varepsilon^j \text{ and } j \neq i \\ -\sum_{i \in \mathscr{L}} \left\{ \alpha_i \frac{M_i}{2} + \frac{1}{c_i} \binom{\alpha_i}{2} \right\} & \text{if } \beta = \alpha \\ 0 & \text{otherwise .} \end{cases} \tag{1.16}$$

The Q-matrix can be interpreted in an intuitive way: when tracing back the genealogy of the sample, any pair of lineages inside subpopulation i has coalescence rate $1/c_i$, and a particular lineage moves from subpopulation i to subpopulation j at rate $M_{ij}/2$.

The structured coalescence has been used as a framework for modelling demographic scenarios on structured populations cite (Wakeley, others)(Wakeley, 2001). It has been rigorously proved (Herbots, 1994) that, under some hypothesis for reproduction and migration, the genealogy of a sample from a subdivided population in a discrete-time model, after a change in the time scale, is actually the structured coalescent. In chapter 4 we propose a model based on the structured coalescent for studying the evolution of a subdivided population, taking into account the changes in migration rates and population size.

1.4 Methods for demographic inference

In a neutral model of mutations (mutations that do not reduce or increase the chances of individuals to reproduce), the genealogical process is not affected by mutations. As a consequence, selectively neutral mutations can be easily incorporated to the model, based on the idea that the mutation process is independent of the genealogical process. Thus, we can study the joint effects of mutation and genetic drift on genomic data by simulating a genealogy backward in time and then adding mutations on the tree according to a Poisson process (Hudson, 2002). This is equivalent (and by far more efficient) that simulating the entire population for a long number of generations forward in time. The idea that the observed patterns on present DNA sequences are the result of random mutations on a random tree has completely changed the way we see genetic data. This has been a fundamental contribution of the coalescent, beside its mathematical interest.

Key events taking place in the evolutionary process can leave marks on the genealogies of individuals sampled at present times. Hence, methods allowing to figure out the shape genealogies have, could throw some light on the history of populations. On the other hand, even if the neutral mutation process and the genealogical process can be studied independently, the patterns of mutations on present data are strongly related with the genealogy. Neutral mutations, by definition, don't affect the genealogy, but the way the mutation process is modeled (a Poisson process over the given genealogy) makes that mutations on the sample are strongly dependent on the underlying genealogy. For example, if the ancestor of two genes is far in the past (corresponding with a higher branch length), it is more likely to have mutations over the branches of the genealogical tree. These mutations we will be reflected on the data by differences between genes at some positions in the DNA sequences. Following this reasoning, for a given genealogy

and a fixed mutation rate, we can compute the probability to observe a particular pattern of mutations in the present sample. In other words, we can compute the *likelihood* of any given genealogy with respect to the observed pattern of mutations. No matter how big a sample is, the underlying genealogy is unique (the genealogical tree of the ancestors from the present to the MRCA) and, in principle, it could be estimated using a maximum likelihood strategy. However this is impossible in practice due to the high dimension of the space of all possible genealogies. Given that the underlying genealogy is unknown in most of the cases, the methods used to reconstruct the demographic history sometimes take the integral (or rather an approximation) over all possible underlying genealogies.

The coalescent theory along with the subsequent theoretical developments allow to describe the relationship between population size over time and the genealogy of genes. This relationship has been exploited by many coalescent-based methods with the aim of inferring the demographic history from DNA sequences in different ways:

- Compare observed distribution of *pairwise genetic differences* with expected distributions derived from coalescence theory
- Given a specific model for sequence evolution and assumed some deterministic demographic changes (for example constant population size, exponential grow, bottleneck), compute the likelihood of an observed set of DNA sequences.
- Infer past demographic history from a reconstructed genealogy.

In Beaumont (1999) a method is presented for detecting expansions or declines of a population. Going back in time, it is considered that the population size changes from a value N_0 (which is size at the time when the sample has been taken, and considered as the present population size) to an ancestral value N_1 and stay constant thereafter. Forward in time, this corresponds to a population that was constant in the past until some time (denoted t_f) and then began to change in size from t_f until the present. Two demographic models are considered: *linear* population size change with ratio $r = \frac{N_0}{N_1}$ or *exponential* population size change with rate r. For each model, assuming the underlying genealogy is known, the likelihood of the parameters r and t_f with respect to the observed data is described based on Griffiths and Tavaré (1994). Then, the integral over all possible genealogies is approximated by Markov Chain Monte Carlo (MCMC) simulations. This makes it possible to find approximations to the posterior Likelihood of the parameters.

A method for inferring demographic history based on gene genealogies was introduced by Pybus et al. (2000). Given a set of DNA sequences, the authors

reconstruct more plausible genealogy from a set G^* of possible genealogies, by a Maximum Likelihood approach as proposed by Felsenstein (1981). Then, based on the estimated genealogy, a Maximum Likelihood Estimation strategy allows to select, from a set of candidate demographic scenarios H, the one that maximise the likelihood with respect to the reconstructed genealogy. The hypothetical scenarios are piecewise constant functions that authors call *skyline plots* and represent the population size at different time intervals. Unlike other approaches developed until then, the method of Pybus et al. (2000) does not assume any prior demographic scenario (many other methods assume a particular function of population size changes, ex: linear grow, constant, bottleneck). For this reason, authors stated that the framework is a nonparametric way to do estimates. However, the method is still based on the coalescence theory for changing population size (Griffiths and Tavaré, 1994) which assumes that population is panmictic.

The spectacular progress of genotyping and sequencing technologies during the last decade has enabled the production of high density genome-wide data in many species. New statistical methods accounting for recombination and scalable to the analysis of whole genome sequences have been proposed. Some of them (Li and Durbin, 2011; Schiffels and Durbin, 2014; Sheehan et al., 2013) are based on the Sequentially Markovian Coalescent model (McVean and Cardin, 2005a; Marjoram and Wall, 2006), an approximation of the classical coalescent with recombination (Hudson, 1983), where coalescent trees are assumed to be Markovian along the genome. Thanks to this Markovian assumption, maximum likelihood estimates of past population sizes can be efficiently obtained from the observation of one or several diploid genomes.

Other statistical methods for estimating population size history are based on the ABC approach (Beaumont et al., 2002; Csilléry et al., 2010; Beaumont, 2010). For example, PopSizeABC (Boitard et al., 2016) estimates complex population size histories involving many population size changes from a sample of whole-genome sequences. To do this, the method considers two classes of summary statistics which are very informative about past population size: the folded allele frequency spectrum (AFS) and the average linkage disequilibrium (LD) at different physical distances. Combining these summary statistics with the ABC framework allows to do accurate estimations of the population sizes from the first few generations before the present back to the expected time to the most recent common ancestor of the sample. Details about this method can be found in Annexes.

Most of the methods proposed for estimating demographic history are based on the assumption that population is panmictic. As we will see in the next section, this may be problematic when we analyse a population which is structured.

1.5 Confounding effects of population size changes in structured population

Whereas methods to infer population size changes have become increasingly popular a growing number of studies (Wakeley, 1999; Vogl et al., 2003; Städler et al., 2009; Chikhi et al., 2010; Heller et al., 2013) have found that when populations are structured spurious population size changes can be detected. For instance, in Chikhi et al. (2010), the authors found that the method proposed in Beaumont (1999) for detecting and quantifying population size changes using microsatellites was sensitive to the effects of structure. The analysis was done using the method MSVAR (Beaumont, 1999) applied to simulated data under the n-island and stepping-stone models. It was shown that MSVAR inferred population size changes even though data were simulated assuming a constant size population. Moreover, the results of the parameter estimations were different under the same scenario for different values of migration or different sampling scheme. In another study (Heller et al., 2013) also found signals of recent population decrease when the Bayesian Skyline Plot method (Drummond et al., 2005) was applied to data simulated under a structured scenario. The reconstructed demographic history was also sensitive to the sampling scheme in the scenarios analysed in Heller et al. (2013).

Following the work of these authors and others we explored the effects of the structure on the estimation of population size changes as inferred by the method implemented in PopSizeABC (Boitard et al., 2016). PopSizeABC is based on an ABC approach and uses information from linkage disequilibrium (LD) and genomic data from several individuals to infer a history of population size changes similar to those inferred by the PSMC or MSMC. We simulated data under two scenarios involving an n-island model with ten islands and migration rate of one. In both scenarios the size of one island was considered to be 500, meaning that the size of the entire population was equal to 5000 (the size of a single deme as well as that of the metapopulation are represented with dotted horizontal lines in Figures 1.7 and 1.8). The population size was assumed constant in both cases.

The first scenario consider a simple n-island model. We simulated 50 haploid sequences sampled from the same island and applied PopSizeABC. The method detected a decrease on the population size: starting about $10,000$ generations before the present, the population size decreased from a value of $10,000$ to a value close to 1000 at about 100 generations before the present (Figure 1.7, left panel). We then changed the sampling scheme and simulated 50 haploid sequences, sampling 5 haploid sequences in each island. We found a substantially different history indicating a population size close to 5000 at about $100,000$ generations before the present, followed by an increasing between $100,000$ generations and 8000

generations before the present and then a decrease from 8000 to 500 generations before the present, staying close to an effective size of 5000 from 500 generations before the present (Figure 1.7, right panel).

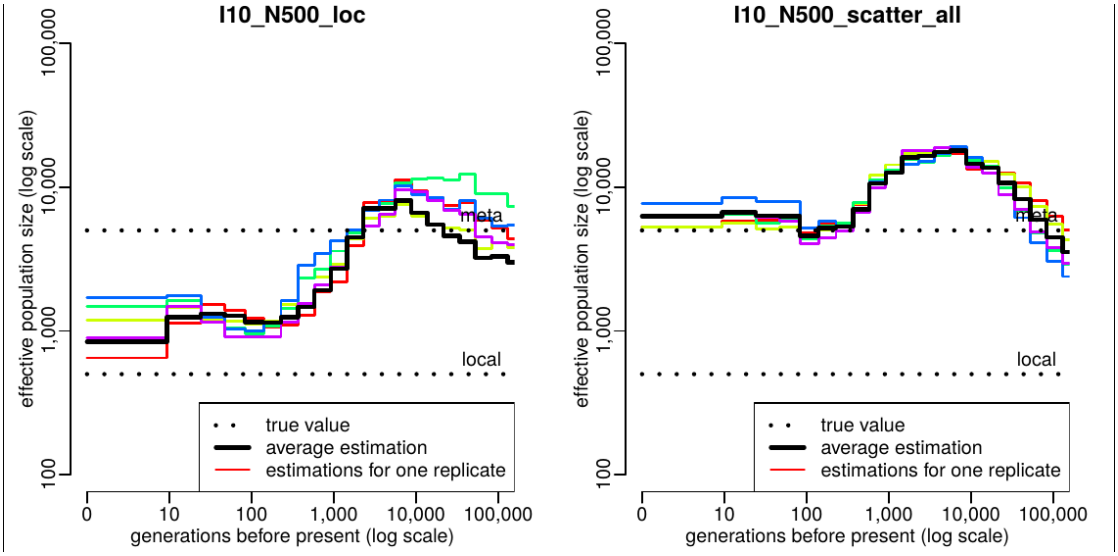


Figure 1.7: Effects of the structure on the estimation of PopSizeABC with different sample schemes.

The second scenario considers that the 10 subpopulations of effective size 500 diverged from a population that was panmictic with a population size of 5000, according to an isolation with migration model (Nielsen and Wakeley, 2001). The migration rate at the present is supposed symmetrical and equal to one. The PopSizeABC analysis was done from 50 haploid sequences sampled from the same island. The divergence occurred 40 generations before the present (Figure 1.8 left panel) and 200 generations before the present (Figure 1.8 right panel). We found in both cases a decreases from a population size close to 7000, 1000 generations before the present, to a population size close to 600, 50 generations before the present.

As we can see, even complex and recent methods for inferring the population size that uses LD patterns can be sensitive to population structure. A theory should thus be developed in order to decide if the changes detected by these methods are related with past population size changes or are mainly effects of population structure, and changes thereof.

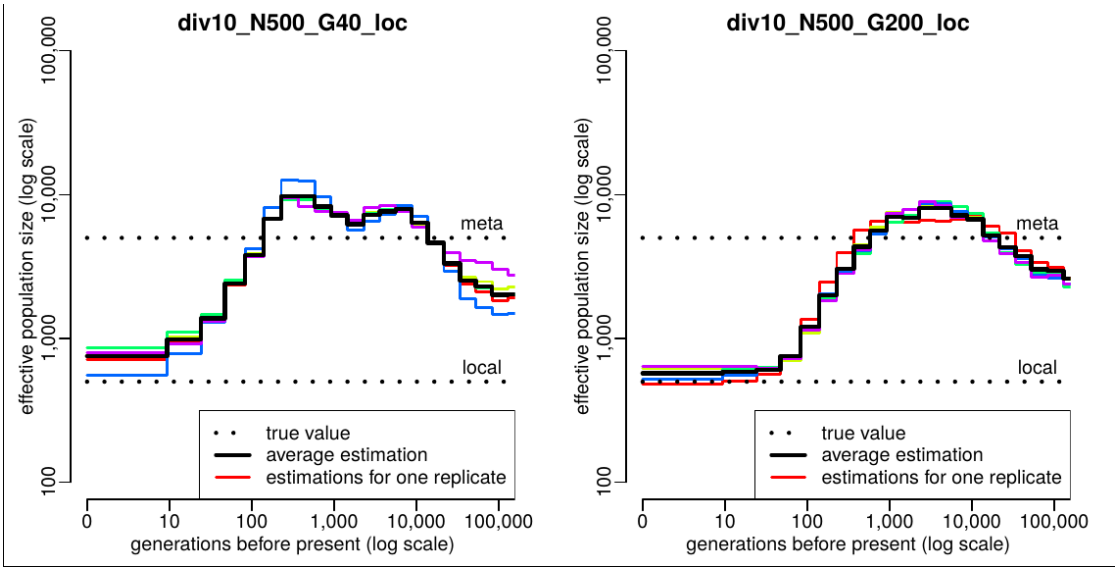


Figure 1.8: Estimation of PopSizeABC when applied to a scenario of isolation with migration.

Chapter 2

Demographic inference using genetic data from a single individual: separating population size variation from population structure

In section 1.5 we illustrated how methods assuming that population is panmictic tend to infer a demographic history with signals of population size change when applied to a population which is structured, even though the total population size remains constant. This makes us question whether the history inferred using many methods currently available actually identifies real population size changes. A way to address this would be to use a method allowing to determine whether the population under study is closer to a panmictic model than to a structured model. In the following chapter we present our first attempt at addressing this complex issue. We started by studying the properties of the distribution of coalescent times under two simple demographic models. This allowed us to see that even though structured models generated signals similar to bottleneck populations the statistical distributions were different. As a consequence this allowed us to develop a method aiming to distinguish a structured population from a panmictic population subjected to a single population size. This chapter describes this work and it is nearly identical to the published work Mazet et al. (2015b). The only difference is that we have added at the end (section 2.7) some theoretical proof and preliminary results on the application to real data which were not done when the study was submitted and published.

Abstract

The rapid development of sequencing technologies represents new opportunities for population genetics research. It is expected that genomic data will increase our ability to reconstruct the history of populations. While this increase in genetic information will likely help biologists and anthropologists to reconstruct the demographic history of populations, it also represents new challenges. Recent work has shown that structured populations generate signals of population size change. As a consequence it is often difficult to determine whether demographic events such as expansions or contractions (bottlenecks) inferred from genetic data are real or due to the fact that populations are structured in nature. Given that few inferential methods allow us to account for that structure, and that genomic data will necessarily increase the precision of parameter estimates, it is important to develop new approaches. In the present study we analyse two demographic models. The first is a model of instantaneous population size change whereas the second is the classical symmetric island model. We (i) re-derive the distribution of coalescence times under the two models for a sample of size two, (ii) use a maximum likelihood approach to estimate the parameters of these models (iii) validate this estimation procedure under a wide array of parameter combinations, (iv) implement and validate a model rejection procedure by using a Kolmogorov-Smirnov test, and a model choice procedure based on the AIC, (v) derive the explicit distribution for the number of differences between two non-recombining sequences. Altogether we show that it is possible to estimate parameters under several models and perform efficient model choice using genetic data from a single diploid individual.

2.1 Introduction

The sheer amount of genomic data that is becoming available for many organisms with the rapid development of sequencing technologies represents new opportunities for population genetics research. It is hoped that genomic data will increase our ability to reconstruct the history of populations (Li and Durbin, 2011; Schiffels and Durbin, 2014) and detect, identify and quantify selection (Vitti et al., 2013). While this increase in genetic information will likely help biologists and anthropologists to reconstruct the demographic history of populations, it also exposes old challenges in the field of population genetics. In particular, it becomes increasingly necessary to understand how genetic data observed in present-day populations are influenced by a variety of factors such as population size changes, population structure and gene flow (Nielsen and Beaumont, 2009). Indeed, the use of genomic data does not necessarily lead to an improvement of statistical inference. If the model assumed to make statistical inference is fundamentally mis-specified, then increas-

ing the amount of data will lead to increased precision for perhaps misleading if not meaningless parameters and will not reveal new insights (Nielsen and Beaumont, 2009; Chikhi et al., 2010; Heller et al., 2013).

For instance, several recent studies have shown that the genealogy of genes sampled from a deme in an island model is similar to that of genes sampled from a non structured isolated population submitted to a demographic bottleneck (Chikhi et al., 2010; Heller et al., 2013). As a consequence, using a model of population size change for a spatially structured population may falsely lead to the inference of major population size changes (Nielsen and Beaumont, 2009; Städler et al., 2009; Chikhi et al., 2010; Heller et al., 2013; Paz-Vinas et al., 2013). Conversely, assuming a structured model to estimate rates of gene flow when a population has been submitted to a population size change, may also generate misleading conclusions, even though the latter case has been much less documented. More generally, previous studies have shown that spatial processes can mimic selection (Currat et al., 2006), population size changes (Leblois et al., 2006; Chikhi et al., 2010; Heller et al., 2013) or that changes in gene flow patterns can mimic changes in population size (Wakeley, 1999; Broquet et al., 2010). The fact that such dissimilar processes can generate similar coalescent trees poses exciting challenges (Nielsen and Beaumont, 2009). One key issue here is that it may be crucial to identify the kind of model (or family of models) that should be used before estimating and interpreting parameters.

One solution to this problem is to identify the "best" model among a set of competing models. This research program has been facilitated by the development of approximate Bayesian computation (ABC) methods (Beaumont et al., 2002; Cornuet et al., 2008; Beaumont, 2010). For instance, using an ABC approach, Peter et al. (2010) showed that data sets produced under population structure can be discriminated from those produced under a population size change by using up to two hundred microsatellite loci genotyped for 25 individuals. In some cases, relatively few loci may be sufficient to identify the most likely model (Sousa et al., 2012; Peter et al., 2010), but in others, tens or hundreds of loci may be necessary (Peter et al., 2010). ABC approaches are thus potentially very powerful but they are often used as black boxes which provide results on a specific problem but limited understanding on the properties of genetic data in general. Also, since most ABC methods use summary statistics, which are rarely sufficient they typically lose part of the information present in the genetic data compared to likelihood-based methods (Beaumont, 2010). Analytical approaches on the contrary are often limited to very simple models and do not exhibit the flexibility of ABC methods but they allow us to improve our understanding of genetic data. For instance, the theory developed for the coalescent under structured models is crucial to understand why population structure mimics population size changes. Below, we use intuitive and

analytical results to explain exactly that and identify connections between models and parameters that would typically be missed with ABC approaches.

In the present study we are interested in describing the properties of the coalescent under two demographic models and in devising a new statistical test and new parameters estimation procedures. The two models were a model of population size change and a model of population structure. More specifically we re-derived the full distribution of T_2, the time to the most recent common ancestor for a sample of size two for a model of sudden population size change and for the *n-island* model. We then used a maximum likelihood-like approach to estimate the parameters of interest for each model (timing and ratio of population size change for the former and number of migrants and number of islands for the latter). We developed a statistical test that identifies data sets generated under the two models and an AIC (Akaike Information Criterion) model choice procedure for the cases where both models were rejected. We also tested the robustness of our model choice approach by simulating data under four other models, two models of population size change and two stepping-stone models. Finally, we show how these results may apply to genomic data such as SNPs and how they could be extended to real data sets (for which the T_2 is not usually known) and for other demographic models. In particular we discuss how our results are relevant in the context of the PSMC (Pairwise Sequentially Markovian Coalescent) method (Li and Durbin, 2011), which has been now extensively used on genomic data and also uses a sample size of two.

2.2 Demographic models

2.2.1 Population size change:

We consider a simple model of population size change, where $N(t)$ represents the population size (N, in units of genes or haploid genomes) as a function of time (t) expressed in generations scaled by N, the population size, and where $t = 0$ is the present, and positive values represent the past (Figure 2.1 (a)). More specifically we assume a sudden change in population size at time T in the past, where N changes instantaneously by a factor α. This can be summarized as $N(t) = N(0) = N_0$ for $t \in [0, T[$, $N(t) = N(T) = \alpha N_0$ for $t \in [T, +\infty[$. If $\alpha > 1$ the population went through a bottleneck (Figure 2.1) whereas if $\alpha < 1$ it expanded. Since N represents the population size in terms of haploid genomes, the number of individuals will therefore be $N/2$ for diploid species. Note also that for a population of constant size the expected coalescence time of two genes is N generations, which therefore corresponds to $t = 1$. In other words, one unit of standardized time corresponds to N generations. We call this model the SSPSC,

which stands for Single Step Population Size Change.

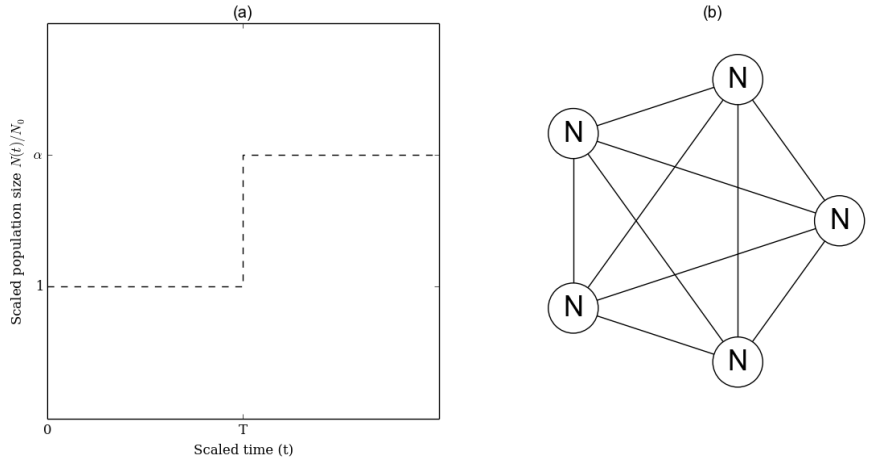


Figure 2.1: Demographic models. (a): Single step population size change (SSPSC) model. The x-axis represents t, the time to the past in units of generations scaled by the number of genes. At time t = T, (going from the present to the past) the population size changes instantaneously from N0 to N1 by a factor α. The y-axis represents the population sizes in units of N_0 (*i.e.* $N(t)/N(0)$). (b): Structured symmetrical island (StSI) model for $n = 5$ islands. Each circle represents a deme of size N. All demes are connected to each other by symmetrical gene flow, represented by the edges. In this example the total number of genes is $5N$. Note that these two models are scaled such that N_0 in the $SSPSC$ model corresponds to N in the StSI model. This implicit scaling is natural since by setting the number of islands to $n = 1$, the two models will be identical for $\alpha = 1$ too, leading to $N_0 = N$.

2.2.2 Structured population:

Here we consider the classical symmetric *n-island* model Wright (1931), see Figure 2.1 (b), where we have a set of n islands (or demes) of constant size N, interconnected by gene flow with a migration rate m, where $\frac{M}{2} = Nm$ is the number of immigrants (genes) in each island every generation. The whole metapopulation size is therefore nN (this is the total number of genes or haploid genomes). Again, N is the number of haploid genomes, and $N/2$ the number of diploid individuals. We call this model the StSI, which stands for Structured Symmetrical Island model.

2.3 The distribution of coalescence times: qualitative and quantitative analyses

In this section we used previous results (Herbots, 1994; Donnelly and Tavaré, 1995) to derive the distribution of coalescent times for the two models of interest. We show that even though they are different, these distributions can be similar under an indefinitely large number of parameter values (Figures 2.2 and 2.3). Moreover we show that even when the distributions are distinguishable, their first moments may not be. In particular, we show that the first two moments (mean and variance) are near identical for a large number of parameter combinations. Before doing that we start by providing a simple intuitive rationale explaining why and how a model of population structure can be mistaken for a model of population size change. This intuitive approach is important because it allows us to understand how the parameters of the two models ((T, α) and (M, n), respectively) are linked.

2.3.1 Intuitive and qualitative rationale:

We start by taking two genes sampled in the present-day population under the Single Step Population Size Change (SSPSC) model. If we assume that $\alpha > 1$ (population bottleneck from an ancient population of size N_1 to a current population of size N_0, with $N_1 = \alpha N_0$) the probability that the two genes coalesce will vary with time as a function of N_0, N_1 and T. If T is very small, then most genes will coalesce at a rate determined by N_1, whereas if T is very large the coalescence rate will be mostly determined by N_0. If we now take two genes sampled from the same island in the Structured Symmetrical Island (StSI) model, we can also see that their coalescence rate will depend on N, the size of the island and on m, the migration rate. If m is very low, the coalescence rate should mostly depend on N. If m is high, the two genes may see their lineages in different islands before they coalesce. As a consequence the coalescence rate will depend on the whole set of islands and therefore on the product nN, where n is the total number of islands.

This intuitive description suggests that there is an intrinsic relationship between T and $1/M$, and between α and n. The reason why structured populations exhibit signals of bottlenecks is because in the recent past the coalescence rate depends on the local island size N, whereas in a more distant past it depends on nN. In other words, it is as if the population size had been reduced by a factor of n. As we will see this rationale is only qualitatively correct, but it suggests that if we want to distinguish them it may be necessary to derive the full distribution of the coalescence times under the two models. We shall denote these coalescence times T_2^{SSPSC} and T_2^{StSI}, respectively.

2.3.2 Derivation of the distribution of coalescence times:

The distribution of T_2^{SSPSC} The generalisation of the coalescent in populations of variable size was first rigorously treated in Donnelly and Tavaré (1995)), and is clearly exposed in Tavaré (2004)). Details of the derivation can be found in the Supplementary Materials. In the case of the $SSPSC$ model, this leads to the following *pdf*

$$f_{T_2}^{SSPSC}(t) = e^{-t}\mathbb{I}_{[0,T[}(t) + \frac{1}{\alpha}e^{-T-\frac{1}{\alpha}(t-T)}\mathbb{I}_{[T,+\infty[}(t), \tag{2.1}$$

where $\mathbb{I}_{[a,b[}(x)$ is the Kronecker index such that

$$\mathbb{I}_{[a,b[}(x) = \begin{cases} 1 & \text{for } x \in [a,b[\\ 0 & \text{otherwise.} \end{cases}$$

The distribution of T_2^{StSI} Herbots (1994) derived the distribution of the coalescence time T_2^{StSI} of two genes for our structured model, see the Supplementary Materials for details and Hudson et al. (1990) and Griffiths (1981) for further reading. If we set $\gamma = \frac{M}{n-1}$ and if Δ is the discriminant of the polynomial D, with $D = \theta^2 + \theta(1+n\gamma) + \gamma$, then the two solutions of D are

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

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

and if we set

$$a = \frac{\gamma - \alpha}{\beta - \alpha} = \frac{1}{2} + \frac{1 + (n-2)\gamma}{2\sqrt{\Delta}}.$$

we then obtain the *pdf* of T_2^{StSI} which is an exponential mixture:

$$f_{T_2}^{StSI}(t) = ae^{-\alpha t} + (1-a)e^{-\beta t}. \tag{2.2}$$

2.3.3 First moments:

Equations 2.4 and 2.5 are different hence showing that it is in principle possible to identify genetic data produced under the two demographic models of interest. The two equations can be used to derive the expectation and variance of the two random variables of interest, T_2^{SSPSC} and T_2^{StSI}. Their analytic values can be easily expressed as functions of the model parameters:

$$
\begin{aligned}
\mathbb{E}\left(T_2^{SSPSC}\right) =& 1 + e^{-T}(\alpha - 1), \\
Var\left(T_2^{SSPSC}\right) =& 1 + 2Te^{-T}(\alpha - 1) + 2\alpha e^{-T}(\alpha - 1) - (\alpha - 1)^2 e^{-2T}, \\
\mathbb{E}(T_2^{StSI}) =& n, \\
Var(T_2^{StSI}) =& n^2 + \frac{2(n-1)^2}{M}.
\end{aligned}
$$

It is interesting to note that the expected time in the StSI model is n and does not depend on the migration rate (Durrett, 2008). The variance is however, and as expected, a function of both n and M. For the SSPSC model, the expected coalescence time is a function of both T and α. We note that it is close to 1 when T is very large and to α when T is close to zero. Indeed, when the population size change is very ancient, even if α is very large the expected coalescence time will mostly depend on the present-day population size, N_0. Similarly, when T is small it will mostly depend on N_1. The relationship that we mentioned above between n and α (and between M and $1/T$) can be seen by noting that when T is close to zero (and M is large), the expectations under the two models are α and n, and the variances are $Var\left(T_2^{SSPSC}\right) \approx 1 + 2\alpha(\alpha - 1) - (\alpha - 1)^2 = \alpha^2$ and $Var(T_2^{StSI}) \approx n^2$. This exemplifies the intuitive rationale presented above. This relationship is approximate and will be explored below, but can be illustrated in more general terms by identifying scenarios with similar moments.

As figure 2.2 shows, the two models provide near-identical pairs of values for $(\mathbb{E}(T_2), Var(T_2))$ for "well chosen" parameters (T, α) and (M, n). Here by setting T to 0.1 (and M to 9, *i.e.* $1/M \approx 0.11$) whereas α and n were allowed to vary from 1 to 100, and from 2 to 100, respectively, we see that the two models exhibit very similar behaviours. We also plotted a second example obtained by setting M to 0.5 and T to 1.09, and varying n and α as above. These examples ilustrate how n and α (respectively, M and $1/T$) are intimately related.

The near-identical values obtained for the expectation and variance under the two models explain why it may be difficult to separate models of population size change from models of population structure when the number of independent genetic markers is limited. However, the differences between the distributions of coalescence times under the two models suggest that we can go further and identify one model from another. For instance, figure 2.3 shows that even in cases where the first two moments are near-identical ($T = 0.1$ and $\alpha = 10$ versus $M = 7$ and $n = 9$), it should be theoretically possible to distinguish them. This is exactly what we aim to do in the next section. In practice, we will assume that we have a sample of n_L independent T_2 values (corresponding to n_L independent *loci*) and will use these T_2 values to (i) estimate the parameter values that best explain

this empirical distribution under the two models of interest, (ii) use a statistical test to compare the empirical distribution with the expected distribution for the maximum likelihood (ML) estimates and reject (or not) one or both of the models. For simplicity, and to make it easier to read, we will often use the term *loci* in the rest of the manuscript when we want to mention the number of independent T_2 values.

2.4 Model choice and parameter estimation

2.4.1 General principle and parameter combinations:

Given a sample $(t_1, ..., t_{n_L})$ of n_L independent observations of the random variable T_2, we propose a parameter estimation procedure and a goodness-of-fit test to determine whether the observed distribution of the T_2 values is significantly different from that expected from the theoretical T_2^{SSPSC} or T_2^{StSI} distributions. This sample can be seen as a set of T_2 values obtained or estimated from n_L independent loci. We took a ML approach to estimate the parameters (T, α) and (M, n) under the hypothesis that the n_L-sample was generated under the T_2^{SSPSC} and the T_2^{StSI} distributions, respectively. We note here that the ML approach was applied to a reduced parameter space due to the fact that the likelihood is actually unbounded (see Supplementary materials for the details of the estimation procedure). The ML estimates $(\widehat{T}, \widehat{\alpha})$ and $(\widehat{M}, \widehat{n})$ were then used to define T_2^{SSPSC} or T_2^{StSI} reference distributions. The Kolmogorov-Smirnov (KS) test which allows to compare a sample with a reference distribution was then used to determine whether the observed n_L sample could have been generated by the respective demographic models. In other words this allowed us to reject (or not) the hypothesis that the $(t_1, ..., t_{n_L})$ sample was a realization of the reference distributions (T_2^{StSI} or T_2^{SSPSC}). Note that the estimation procedure and the KS test were performed on independent sets of T_2 values. We thus simulated twice as many T_2 values as needed ($2n_L$ instead of n_L). With real data that would require that half of the loci be used to estimate $(\widehat{T}, \widehat{\alpha})$ and $(\widehat{M}, \widehat{n})$, whereas the other half would be used to perform the KS test.

We expect that if the estimation procedure is accurate and if the KS test is performing well we should reject the SSPSC (respectively, the StSI) model when the data were simulated under the StSI (resp., the SSPSC) model. On the contrary we should not reject data simulated under the SSPSC (resp., the StSI) model when they were indeed simulated under that model. To validate our approach we used $(t_1, ..., t_{2n_L})$ data sampled from the two T_2 distributions and quantified how the estimation procedure and the KS test performed. In order to do that, we varied the parameter values ((T, α) and (M, n)) for various $2n_L$ values as follows. For T and α we used all 36 pairwise combinations between these two sets of values (0.1,

0.2, 0.5, 1, 2, 5), and (2, 4, 10, 20, 50, 100), respectively. For M and n we used all the 48 combinations between the following values (0.1, 0.2, 0.5, 1, 5, 10, 20, 50) and (2, 4, 10, 20, 50, 100), respectively. For $2n_L$ we used the following values (40, 100, 200, 400, 1000, 2000, 20000). Altogether we tested 588 combinations of parameters and number of loci. For each $2n_L$ value and for each parameter combination (T, α) (or (M, n)) we realized 100 independent repetitions of the following process. We first simulated a sample of $2n_L$ values using the *pdfs* of the SSPSC (resp. StSI) model with (T, α) (resp. (M, n)). We then used the first n_L values to obtain the ML estimates $(\widehat{T}, \widehat{\alpha})$ for the SSPSC model and $(\widehat{M}, \widehat{n})$ for the StSI model. Then, we performed a KS test using a 0.05 threshold on the second half of the simulated data (*i.e.* n_L values) with each of the theoretical distributions defined by the estimated parameters. Finally, after having repeated this process 100 times we recorded all estimated parameters and counted the number of times we rejected the SSPSC and StSI models for each parameter combination and each $2n_L$ value.

2.4.2 Maximum Likelihood Estimation (MLE) in the SSPSC case:

We know from equation 2.4 the *pdf* of the coalescence time in the SSPSC model of two genes. We can thus write the likelihood function for any couple of parameters (α, T), given one observation t_i as:

$$\mathbb{L}_{t_i}(\alpha, T) = \frac{1}{\alpha} e^{-T-\frac{1}{\alpha}(t_i - T)} \mathbb{I}_{[0,t_i[}(T) + e^{-t_i} \mathbb{I}_{]t_i,+\infty[}(T).$$

Given n_L independent values $t = (t_1, t_2, ..., t_{n_L})$, the likelihood is:

$$\mathbb{L}_{SSPSC}(\alpha, T) = \prod_{i=1}^{n_L} \mathbb{L}_{t_i}(T, \alpha),$$

and taking the *log* it gives:

$$\log(\mathbb{L}_{SSPSC}(\alpha, T)) = \sum_{i=1}^{n_L} \log(\mathbb{L}_{t_i}(\alpha, T)).$$

Lemma 1. *Given a set of n_L independent observations $\{t_1, t_2, ..., t_{n_L}\}$, if we restrict the domain of the log-likelihood function $\log(\mathbb{L}_{SSPSC})$ to the set $\{(\alpha, t) \in \mathbf{R}^2 | \alpha > 0,\ t < max_{i \in \{1..n_L\}}(t_i)\}$, all the critical points are of the form*

$$m_a = (\alpha_a, t_a), a \in \{1, 2, ..., n_L\}.$$

with

$$\alpha_a = \frac{1}{K}\sum_{i=1}^{n_L} t_i \mathbb{I}_{t_a \leq t_i} - t_a \qquad and \qquad K = \sum_{i=1}^{n_L} \mathbb{I}_{t_i < t_a}.$$

This lemma means that all the local maxima we are interested in, are located at the points m_a, which are necessarily on the vertical lines of the form $\{(\alpha, t_a), \alpha \in \mathbb{R}^+\}$, $a \in \{1, 2, ..., n_L\}$. The search procedure is thus simplified since we have n_L candidates for approximating the MLE. Amongst those n_L points, we take the one that maximizes the log-likelihood function : $(\hat{\alpha}, \hat{T}) = argmax_{a \in \{1,...,n_L\}}\{\log(\mathbb{L}_{SSPSC}(m_a))\}$. For the proof and some comments, see Supplementary Materials.

2.4.3 MLE in the StSI case:

Under the StSI model the expression of the critical points is not analytically derived. We know from equation 2.5 the *pdf* of coalescence times for two genes. Given n_L independent values $t = (t_1, t_2, ..., t_{n_L})$ we can compute the log-likelihood function for any set of parameters (n, M) as:

$$\log(\mathbb{L}_{StSI}(n, M)) = \sum_{i=1}^{n_L} \log(ae^{-\alpha t_i} + (1-a)e^{-\beta t_i}))$$

We used the Nelder-Mead method (Nelder and Mead, 1965) implementation of *scipy* (Jones et al., 2001) to find numerically an approximation to the maximum of the likelihood function. This method returns a pair of real numbers $(\hat{n}, \hat{M})$. Since n should be an integer we kept either $\lfloor \hat{n} \rfloor$ or $\lfloor \hat{n} \rfloor + 1$, depending on which had the largest log-likelihood value.

2.4.4 Akaike Information Criterion and robustness to model departures:

Once we have computed our aproximations to the MLE for each case (*i.e.* $(\hat{\alpha}, \hat{T})$ for SSPSC and $(\hat{n}, \hat{M})$ for StSI), we proceed to do the KS test. At this stage it is possible to reject both models (or none of them if the data are not sufficiently informative). Rejection of both models may arise as a consequence of various factors such as estimation errors or when the data were produced by models different from the SSPSC and StSI models (see below). By using an *Akaike Information Criterion* (AIC) (Akaike, 1974), it may still be possible to identify which of the two models is the most likely to explain the data. We carried out additional simulations (see Supplementary Materials) to illustrate how the AIC allows us to select the closest model when the KS test rejects the two models even though the data were generated by one of them. Note that our reference models are both characterized by two parameters. Therefore, a simple comparison of the MLE is enough

to make a choice. Nevertheless, the AIC values are easy to compute and they can be useful in order to quantify the information loss when we choose one model rather than the other. The AIC procedure is also more general and could be used to compare more complex models.

Indeed, if the data were generated by different models of population size change or population structure, it would be important to determine whether our approach would allow us to identify the closest model. For instance, if the data were generated by a model of population structure different from the StSI model, the AIC may identify the StSI as the best model even if it is rejected by our KS test. As a test of robustness we carried out additional simulations with data generated under four demographic models departing from our two simplistic models. The first model is analogous to our SSPSC but with four instantaneous population size changes at four different moments in the past. The second one is a model of exponential population size change similar to that of Beaumont (1999), with a recent exponential expansion. The third and fourth are symmetrical stepping-stone models with 16 islands (4×4) and 49 islands (7×7) respectively (Kimura and Weiss, 1964). For consistency we call them 4SPSC (four steps population size change), SEPSC (single exponential population size change), 4x4StSSS and 7x7StSSS (structured symmetrical stepping-stone). For these models the KS is expected to reject both the SSPSC and StSI most of the time, when n_L is large. However, when we apply the AIC procedure we should identify the StSI model as the best model when data were simulated under the two StSSS models, and we should identify the SSPSC as the best model when data were simulated under the 4SPSC and SEPSC. We used the *ms* software (Hudson, 2002) to simulate data and we repeated the experiment 100 times for each value of n_L (the sample size).

2.5 Results

Figure 2.4 shows, for various values of n_L, the results of the estimation of α (panels (a), (c), and (e), for simulations assuming $\alpha = 10$ and $T = (0.1, 1, 2)$, respectively ; see Supplementary Material for the other values) and the estimation of n (panels (b), (d), and (f) for simulations with $n = 10$ and $M = (10, 1, 0.5)$, respectively; see Supplementary Material for the other values, corresponding to 26 figures and 168 panels). The first thing to notice is that both α and n are increasingly well estimated as n_L increases. This is what we expect since n_L represents the amount of information (the number of T_2 values or independent loci.) The second thing to note is that the two parameters are very well estimated when we use $10,000$ values of T_2. This is particularly obvious for n compared to α, probably because n must be an integer, whereas α is allowed to vary continuously. For instance, for most simulations we find the exact n value (without error) as soon as we have more than 1000 loci. However, we should be careful in drawing very general rules. Indeed, when fewer T_2 values are available (*i.e.* fewer independent loci), the estimation precision of both parameters depends also on T and M, respectively. Interestingly, the estimation of α and n are remarkable even when these parameters are small. This means that even "mild" bottlenecks may be very well quantified (see for instance the Supplementary materials for $\alpha = 2$, T values between 0.1 and 1 when we use only 1000 loci). We should also note that when the bottleneck is very old ($T = 5$) the estimation of the parameters is rather poor and only starts to be reasonable and unbiased for $n_L = 10,000$. This is not surprising since the expected T_{MRCA} is 1. Under the SSPSC model most genes will have coalesced by $t = 5$, and should therefore exhibit T_2 values sampled from a stationary population (*i.e.* $\alpha = 1$). As the number of loci increases, a small proportion will not have coalesced yet and will then provide information on α. The expected proportion of genes that have coalesced by $t = T = 5$ is 0.993.

Figure 2.5 shows for various values of n_L the results of the estimation of T (panels (a), (c), and (e), for simulations assuming $T = 0.2$ and $\alpha = (2, 20, 100)$, respectively; see Supplementary Material for the other values) and the estimation of M (panels (b), (d), and (f), for simulations with $M = 20$ and $n = (2, 20, 100)$, respectively; see Supplementary Material for the other values). As expected again, the estimates are getting better as n_L increases. For the values shown here we can see that T, the age of the bottleneck, is very well estimated even when $\alpha = 2$ (for $n_L = 10,000$). In other words, even a limited bottleneck can be very precisely dated. For stronger bottlenecks fewer loci (between 500 and 1000) are needed to still reach a high precision. This is particularly striking given that studies suggest that it is hard to identify bottlenecks with low α values (Girod et al., 2011). Interestingly, the panels (b), (d) and (f) seem to suggest that it may be

more difficult to estimate M than T. As we noted above this observation should be taken with care. Indeed, T and M are not equivalent in the same way as α and n. This is why we chose to represent a value of M such that $M = 1/T$, and why one should be cautious in drawing general conclusions here. Altogether this and the previous figure show that it is possible to estimate with a high precision the parameters of the two models by using only 500 or 1000 loci from a single diploid individual. There are also parameter combinations for which much fewer loci could be sufficient (between 50 and 100).

In Figure 2.6 we show some results of the KS test for the two cases (See the Supplementary Materials for the other parameter combinations). In the left-hand panels ((a), (c), and (e)) the data were simulated under the SSPSC model and we used the StSI model as a reference (*i.e.* we ask whether we can reject the hypothesis that genetic data were generated under a structured model when they were actually generated under a model of population size change). In the right-hand panels ((b), (d) and (f)) the same data were compared using the SSPSC model as reference and we computed how often we rejected them using a 5% rejection threshold. The left-hand panels exhibit several important features. The first is that, with the exception of $T_2 = 5$ we were able to reject the wrong hypothesis in 100% of the cases when we used $10,000$ independent T_2 values.

This shows that our estimation procedure (as we saw above in figures 2.4 and 2.5) and the KS test are very powerful. The second feature is that for $T = 5$, the test performs badly whatever the number of independent loci (at least up to $10,000$). This is expected since the expected T_{MRCA} of two genes is $t = 1$, and 99.3% of the loci will have coalesced by $t = 5$. This means that out of the $10,000$, only *c.a.* 70 loci are actually informative regarding the pre-bottleneck population size. Another important feature of the left-hand panels is that the best results are generally obtained for $T = 1, 0.5$ and 2, whichever the value of α. This is in agreement with Girod et al. (2011) in that very recent population size changes are difficult to detect and quantify. The observation is valid for ancient population size changes as well. The right-hand panels are nearly identical, whichever α value we used (see also Supplementary Materials), and whichever number of T_2 values we use. They all show that the KS test always rejects a rather constant proportion of data sets. This proportion varies between 3 and 15%, with a global average of 8.9%. Altogether our KS test seems to be anti-conservative. This is expected when the quality of estimations is low (which is especially true for low n_L values). Moreover, since the KS test uses a reference distribution based on the estimated rather than the true values, it is expected to reject the hypothesis that simulated data come from a SSPSC (or a StSI) model more often than the value of 5%. Slight differences between estimated and real values of the parameters may raise the global average of rejections. As a test we repeated the KS test by using the

true value and used 1000 independent data sets instead of 100, and found that the tests rejected between 4.5% and 5.5% of the data sets.

Figure 2.7 is similar to Figure 2.6 but the data were simulated under the StSI model and the KS test was performed first using the SSPSC model as a reference ((a), (c), (e)) and then using the StSI model as a reference ((b), (d), (f)). The left-hand panels ((a), (c), and (e)) show results when we ask whether we can reject the hypothesis that genetic data were generated under a population size change model when they were actually generated under a model of population structure. In the right-hand panels ((b), (d), and (f)) we computed how often we rejected the hypothesis that genetic data were generated under the StSI model when they were indeed generated under that model of population structure. Altogether, the left-hand panels suggest that the results are generally best when $M = (0.1, 0.2, 1)$, but that we get very good results for most values of M when we have $10,000$ loci and can reject the SSPSC when the data were actually generated under the StSI model. The right-hand panels show, as in Figure 2.6, that for all the values of n_L and n we reject a rather constant proportion of data sets (between 5 and 10%). Altogether the two previous figures (figures 2.6 and 2.7) show that it is possible to identify the model under which the data were generated by using a single diploid individual.

Figure 2.8 shows the effectiveness of the AIC to identify the best model when the data were generated assuming models of population structure or population size change other than the SSPSC and StSI models. The scenarios we considered were the 4SPSC, SEPSC, 4x4StSSS and 7x7StSSS models presented above. When the data were generated under a model of population size change whether it was the 4SPSC or the SEPSC (left panel) the AIC identifies the SSPSC as the best model, **even for low numbers of loci**. When we simulated data under the two stepping-stone models (4x4StSSS and 7x7StSSS, right panel) the situation was slightly different. The AIC allowed us to select the StSI as the "best" model with great probability for all n_L values larger than 400. We note that these results are also evident when one looks at the log-likelihoods (see Supplementary Materials). When n_L increases the probability with which a population size model explains data generated by a structured model (or vice versa) becomes increasingly low.

Figure 2.9 is divided in four panels showing the relationships between T and M (panels (b) and (d), for various values of α and n) and between α and n (panels (a) and (c), for various values of T and M). In each of the panels we simulated data under a model for specific parameter values represented on the x-axis, and estimated parameters from the other model, and represented the estimated value on the y-axis. Since we were interested in the relationship between parameters (not in the quality of the estimation, see above), we used the largest n_L value and plotted the average of 100 independent estimation procedures. In panel (a)

we simulated a population size change (SSPSC) for various T values (represented each by a different symbol) and several values of α on the x-axis. We then plotted the estimated value of $\widehat{n}$ for each case (*i.e.* when we assume that the data were generated under the StSI model). We find a striking linear relationship between these two parameters conditional on a fixed T value. For instance, a population bottleneck by a factor 50 that happened N_0 generations ago ($T = 1$) is equivalent to a structured population with $\widehat{n} \approx 22$ islands (and $\widehat{M} \approx 0.71$). Panel (c) is similar and shows how data simulated under a structured population generates specific parameters of population bottlenecks. Panels (b) and (d) show the relationship between T and M. We have plotted as a reference the curve corresponding to $y = 1/x$. As noted above and shown on this graph, this relationship is only approximate and depends on the value of α and n. Altogether, this figure exhibits the relationships between the model parameters. They show that the qualitative relationships between α and n, and between T and $1/M$ discussed above are real but only correct up to a correcting factor. Still, this allows us to identify profound relationships between population structure and population size change.

2.6 Discussion

In this study we have analysed the distribution of coalescence times under two simple demographic models. We have shown that even though these demographic models are strikingly different (Figure 1) there is always a way to find parameter values for which both models will have the same first two moments (Figure 2.2). We have also shown that there are intrinsic relationships between the parameters of the two models (Figure 2.9). However, and this is a crucial point, we also showed that the distributions were different and could therefore be distinguished using a single diploid individual. Using these distributions we developed a ML estimation procedure for the parameters of both models $(\widehat{T}, \widehat{\alpha})$ and $(\widehat{M}, \widehat{n})$ and showed that the estimates are accurate, given enough genetic markers. We showed that by applying a simple KS test we were able to identify the model under which specific data sets were generated. In other words, we were able to determine whether a bottleneck signal detected in a particular data set could actually be caused by population structure using genetic data from a single individual. We also implemented an AIC procedure to identify the "best" of our two models in cases where the KS test rejected both the SSPSC and StSI models. The AIC approach was tested with the two reference models and with four additional scenarios. Our results suggest that it is thus possible to use our approach to determine whether the population under study is structured or not even when the data were not generated by one of our two models.

The fact that a single individual provides enough information to estimate de-

mographic parameters is in itself striking (see in particular the landmark paper by Li and Durbin (2011)), but the fact that one individual (or rather sometimes as few as 500 or 1000 loci from that one individual) potentially provides us with the ability to identify the best of two (or more) models is remarkable as well. The PSMC (pairwise sequentially Markovian coalescent) method developed by Li and Durbin (2011) reconstructs a theoretical demographic history characterized by population size changes, assuming a single non structured population. Our study does not estimate as many parameters as the PSMC and is currently not applicable to real data (but see below). However, it provides a proof of concept and goes therefore one step further. It is a first step towards a more realistic and perhaps critical reconstruction of the demographic history of populations. The models used here are necessarily simplistic, and several authors have noted that real populations are likely to have gone through complex histories which would require models putting together the two families of scenarios proposed (*i.e.* population structure and population size change). In Wakeley (1999), a model considering a structured population that went through a bottleneck in the past was developed. Wakeley (1999) discussed the idea that, in structured populations and under some conditions, an affective size can be computed which will therefore change when changes in the migration rate or the size of islands (*demes*) occur. He noted that changes in population structure can thus be mistaken for changes in effective population size. This idea is of course older and can be found implicitly or explicitly in studies aiming at computing the effective population size of structured populations (*e.g.* Nei and Takahata (1993)) since the various formulae derived to compute the effective size are functions the migration rate, the number of demes and the deme size. The framework presented here should thus be helpful to the aim of setting these two scenarios apart in order to detect (for example) false bottleneck signals. Nevertheless, while our study provides several new results, there are still several important issues that need to be discussed and much progress that can still be made.

2.6.1 T_2 and molecular data

The first thing to note is that we assume, throughout our study, that we have access to the coalescence times T_2. In real data sets, this is never the case and the T_2 are rarely estimated from molecular data. While this is a limitation, we note that the PSMC actually estimates the distribution of T_2 values. In its default implementation the PSMC software does not output this distribution but it can be modified to do it by using specific commands. The PSMC will then provide a discretized distribution in the form of a histogram with classes defined by the number of time periods for which population size estimates are computed. In any case, this suggests that it is in theory possible to use the theoretical work of Li and

Durbin to generate T_2 distributions, which could then be used with our general approach, to compare the history reconstructed by the PSMC with the $StSI$ model. Moreover, it is possible to use the theory developed here to compute, conditional on the T_2 distribution, the distribution of several measures of molecular polymorphism. For instance, consider an infinite site mutation model with mutation rate θ. Assuming that the coalescent time of two non recombining DNA sequences is t, the number of mutations between them will follow a Poisson distribution with parameter $2t\theta$. This allows us to compute the conditional distribution of N_d, the number of differences between pairs of non recombining sequences as:

$$\mathbb{P}(N_d = k|T_2 = t) = e^{-2t\theta}\frac{(2t\theta)^k}{k!}$$

If we know the density of T_2, it is then possible to compute the distribution of N_d by taking the integral over all possible values of t:

$$\mathbb{P}(N_d = k) = \int_0^{+\infty} \mathbb{P}(N_d = k|T_2 = t) f_{T_2}(t)\, dt$$

by doing the computations for the two models studied here (see details in Supplementary Materials) we get:

For the SSPSC model,

$$\mathbb{P}(N_d^{SSPSC} = k) = \frac{(2\theta)^k}{(2\theta+1)^{k+1}} + (2\theta)^k S_k$$

with

$$S_k = \sum_{i=0}^{k} \frac{e^{-T(2\theta+1)}T^{k-i}}{(k-i)!}\left(\frac{1}{\alpha(2\theta+\frac{1}{\alpha})^{i+1}} - \frac{1}{(2\theta+1)^{i+1}}\right)$$

For the StSI model,

$$\mathbb{P}(N_d^{StSI} = k) = \frac{a}{\alpha+2\theta}\left(\frac{1}{1+\frac{\alpha}{2\theta}}\right)^k + \frac{1-a}{\beta+2\theta}\left(\frac{1}{1+\frac{\beta}{2\theta}}\right)^k$$

Applying this to real data and validating it across the parameter space is an important issue that would deserve a full and independent study, which we plan to carry out in the near future.

2.6.2 Error in estimating T_2

As noted in the previous section, we have been assuming that the T_2 values were known without error. As an additional validation step we carried out simulations in which the T_2 values were known with some random error. We considered the

case where T_2 values were estimated with a random noise drawn from a normal distribution with the following standard errors, 1% and 5% (See Supplementary Materials for details). We then used the corresponding T_2 distributions with various n_L values to infer the model parameters and apply the model choice procedures. Our results suggest that even with a standard error of 5% the parameters are well estimated and the model choice procedure is also very efficient. For instance, we identify the right model with 100% success for the chosen parameters with less than $10,000$ loci. As expected the number of loci required to reach a particular level of precision (as measure by the mean standard error, MSE) is larger when the T_2 are estimated with error rather than without error. It is interesting to note that the MSE values seem to reach a plateau for some parameters (α and T) for n_L values between $10,000$ and $100,000$ but not for others (n and M). Altogether, this suggests that even with errors in the estimation of T_2 values a number of loci between $1,000$ and $10,000$ will be enough to estimate the models parameters and to identify or reject models with great confidence.

2.6.3 Demographic models

In our study we limited ourselves to two simple models. It would thus be important to determine the extent to which our approach could be applied to other demographic models. The n-island or StSI model is a classical model whose strongest assumptions is probably that migration is identical between all demes. This is likely to be problematic for species with limited vagility. In fact, for many species a model where migration occurs between neighbouring populations such as the stepping-stone is probably more likely. At this stage it is unclear whether one could derive analytically the *pdf* of T_2 for a stepping-stone model. The work by Herbots (1994) suggests that it may be possible to compute it numerically by inversing the Laplace transform derived by this author. This has not been done to our knowledge. Interestingly, this author has also shown that it is in principle possible to derive analytically the *pdf* of T_2 in the case of a two-island model with populations of different sizes. Again, this would provide us with other structured models against which population size change models could be compared.

The SSPSC model has also been used for several decades (Rogers and Harpending, 1992) and represents a first step towards using more complex models of stepwise population size changes (McManus et al., 2015), or models with more complex trajectories. For instance, the method of Beaumont (1999) to detect, date and quantify population size changes (Goossens et al., 2006; Olivieri et al., 2008; Quéméré et al., 2012; Salmona et al., 2012) assumes either an exponential or a linear population size change. It should be straightforward to compute the *pdf* of T_2 under these two models because the coalescent theory has been very well developed for populations with variable size (Donnelly and Tavaré, 1995; Tavaré,

2004) and it is possible to write the *pdf* of T_2 for any demographic history involving any type of population size changes. Significant work would be needed to apply the general framework outlined here to additional demographic models. But the possibilities opened by this study are rather wide.

2.6.4 Comparison with previous work and generality our of results

The present work is part of a set of studies aimed at understanding how population structure can be mistaken for population size change and at determining whether studies identifying population size change are misleading or valid (Chikhi et al., 2010; Heller et al., 2013; Paz-Vinas et al., 2013). It is also part of a wider set of studies that have recognised in the last decade the importance of population structure as potential factor biasing inference of demographic (Leblois et al., 2006; Städler et al., 2009; Peter et al., 2010; Chikhi et al., 2010; Heller et al., 2013; Paz-Vinas et al., 2013) or selective processes (Currat et al., 2006; Hallatschek and Fisher, 2014). Here we demonstrated that it is possible to separate the SSPSC and StSI models using only one individual. Without undermining this result, we also want to stress that we should be cautious before extending these results to any set of models, particularly given that we only use the information from T_2. Much work is still needed to devise new tests and estimation procedures for a wider set of demographic models and using more genomic information, including recombination patterns as in the PSMC method (Li and Durbin, 2011). Beyond the general approach outlined here we would like to mention the study of Peter et al. (2010) who also managed to separate one structure and one PSC (*Population Size Change*) model. These authors used an ABC approach to separate a model of exponential PSC from a model of population structure similar to the StSI model. Their structured model differs from ours by the fact that it is not an equilibrium model. They assumed that the population was behaving like an n-island model in the recent past, until T generations in the past, but that before that time, the ancestral population from which all the demes of size N derived was not structured and was of size N. When T is very large their model is identical to the StSI, but otherwise it may be quite different. For instance, the fact that their model assumed that the number of demes was 100 means that they also simulated an instantaneous 100-fold population size increase. It is unclear whether such a scenario is necessarily more realistic than Wright's n-island model. Still, the fact that they managed to separate the two models using an ABC approach is promising as it suggests that there is indeed information in the genetic data for models beyond those that we studied here. We can therefore expect that our approach may be applied to a wider set of models. We also stress that these authors used a much

larger sample size (25 diploid individuals corresponding to 50 genes). They used a maximum of 200 microsatellites which corresponds therefore to 10,000 genotypes, a number very close to the maximum number used here. This stresses the complementarity of analytical and ABC approaches. Our study provided new results and several intuitive insights into the relationships of structured and population size change models. We believe that such intuitions would not have been easily found with an ABC approach because ABC methods are often used as black boxes providing results on specific models, rather than general results. For instance we identified the linear relationships between the parameters (α and n, and T and $1/M$). Altogether these analytical developments open up new avenues of research for the distribution of coalescent times under complex models and for larger sample sizes.

2.6.5 Sampling and population expansions

Recent years have also seen an increasing recognition of the fact that the sampling scheme together with population structure may significantly influence demographic inference (Wakeley, 1999; Städler et al., 2009; Chikhi et al., 2010; Quéméré et al., 2012; Heller et al., 2013; Paz-Vinas et al., 2013). For instance, in the n-island model, and under a number of simplifying assumptions (strong migration assumption for instance) genes sampled in different demes will exhibit a genealogical tree similar to that expected under a stationary Wright-Fisher model (Wakeley, 1999). Since our work was focused on T_2 we mostly presented our results under the assumption that the two genes of interest were sampled in the same deme. For diploids this is of course a most reasonable assumption. However, the analytical results presented above also allow us to express the distribution of T_2 when the genes are sampled in different demes. We did not explore this issue further here, but it would be important to study the results under such conditions. Interestingly, we find that if we assume that the two genes are sampled in two distinct demes, we detect population expansions rather than bottlenecks. This could happen if we considered a diploid individual whose parents came from different demes. In that case, considering the two genes sampled in the deme where the individual was sampled would be similar to sampling his two parental genes in two different demes. Interestingly, Peter et al. (2010) noted that when the 25 individuals were sampled in different demes, they would detect population size expansions rather than bottlenecks. This is different from our results since they considered that pairs of alleles would still be in the same deme (since they considered diploids). Our results are therefore complementary and qualitatively in agreement with theirs. Similarly, Heller et al. (2013) also found and noted that signals of population expansion could be detected under scattered sampling schemes. Also, Paz-Vinas et al. (2013) noted that signals of population expansion could be detected in cases

where the sampling scheme changed and when there was asymetrical gene flow between populations.

2.6.6 Conclusion: islands within individuals

To conclude, our results provide a general framework that can be extended to whole families of models. We showed for the first time that genomic data from a single individual can be used to estimate parameters that have to our knowledge never been estimated. During the last decade there has been a major effort to use programs such as STRUCTURE (Pritchard et al., 2000) to estimate the number of "subpopulations" or genetic clusters on the basis of a large number of samples, across the geographical distribution of a particular species. Our work suggests that we can in principle provide additional results and insights with only one individual. It is important to stress though that the answer provided here is different from that obtained with STRUCTURE and similar methods and programs (Pritchard et al., 2000; Guillot et al., 2005; Chen et al., 2007; Corander et al., 2004). We do not aim at identifying the populations from which a set of individuals comes. Rather we show that the genome of a single individual informs us on the whole set of populations, hence including individuals which have not been sampled. In other words, even though we assume that there are n populations linked by gene flow, we show that each individual, is a genomic patchwork from this metapopulation. We find these results reassuring, in an era where genomic data are used to confine individuals to genetic clusters and where division rather than connectivity is stressed.

Beyond this crucial change in outlook towards genomic data, we wish to stress that it is remarkable that we were able to estimate the number of islands (and the number of migrants) in the StSI model. This means that one can in principle use genomic data from non model or model organisms to determine how many islands make up the metapopulation from which one single individual was sampled, and estimate how connected these demes are. This is particularly meaningful for species for which the number of individuals with genomic data is limited. Our ability to estimate n is one of the most striking and powerful results of our study. The number of islands should be obtained across species and individuals for comparative analyses. These results would provide unique insights into the structure of species for which it is difficult to obtain samples in the field such as endangered lemurs (Olivieri et al., 2008; Quéméré et al., 2012).

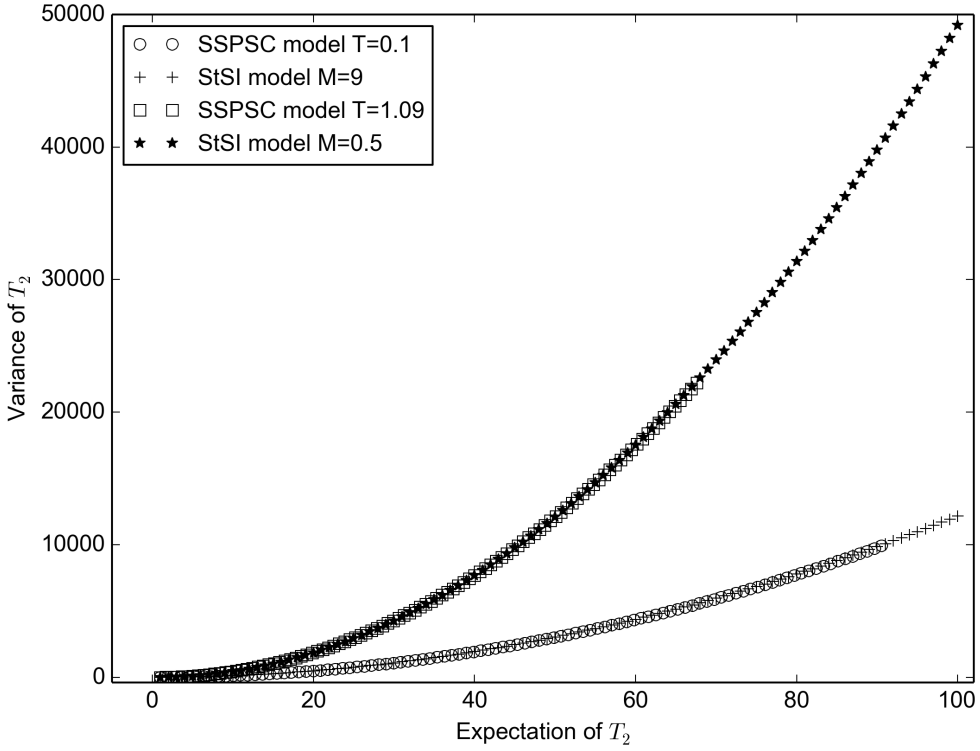


Figure 2.2: Expected value and Variance of T_2 under the SSPSC and StSI models. This figure ilustrates how both models can have the same pair of values $(E(T_2), Var(T_2))$ for many sets of parameters. For the SSPSC model the time at which the population size change occured was fixed to $T = 0.1$ whereas α varied from 1 to 100 in one case, and $T = 1.09$, whereas α varied from 1 to 200 in the other case. For the StSI model the migration rate was fixed to $M = 9$ and $M = 0.5$, whereas n varies from 2 to 100.

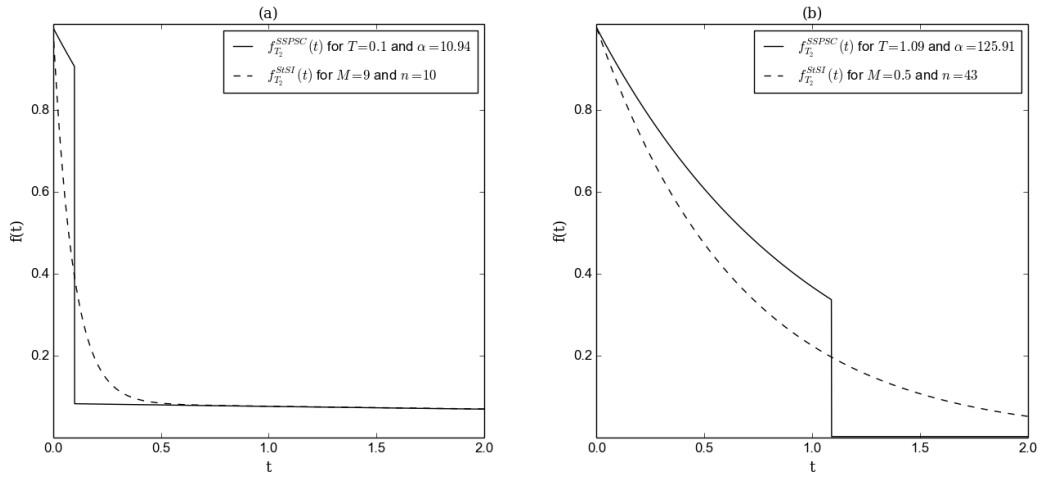


Figure 2.3: Density of T_2 under the SSPSC and StSI models. Two sets of parameter values (panels (a) and (b), respectively) were chosen on the basis that expectations and variances were close. Panel (a): Density for the SSPSC model with $T = 0.1$ and $\alpha = 10.94$, and for the StSI model with $M = 9$ and $n = 10$. For this set of parameters we have $E\left(T_2^{SSPSC}\right) = 9.994$, and $E\left(T_2^{StSI}\right) = 10$, $Var\left(T_2^{SSPSC}\right) = 118.7$ and $Var\left(T_2^{StSI}\right) = 118.0$. Panel (b): The same, but for $T = 1.09$ and $\alpha = 125.91$, and for $M = 0.5$ and $n = 43$. The corresponding expectations and variances are $E\left(T_2^{SSPSC}\right) = 42.997$, and $E\left(T_2^{StSI}\right) = 43$, $Var\left(T_2^{SSPSC}\right) = 8905$ and $Var\left(T_2^{StSI}\right) = 8905$.

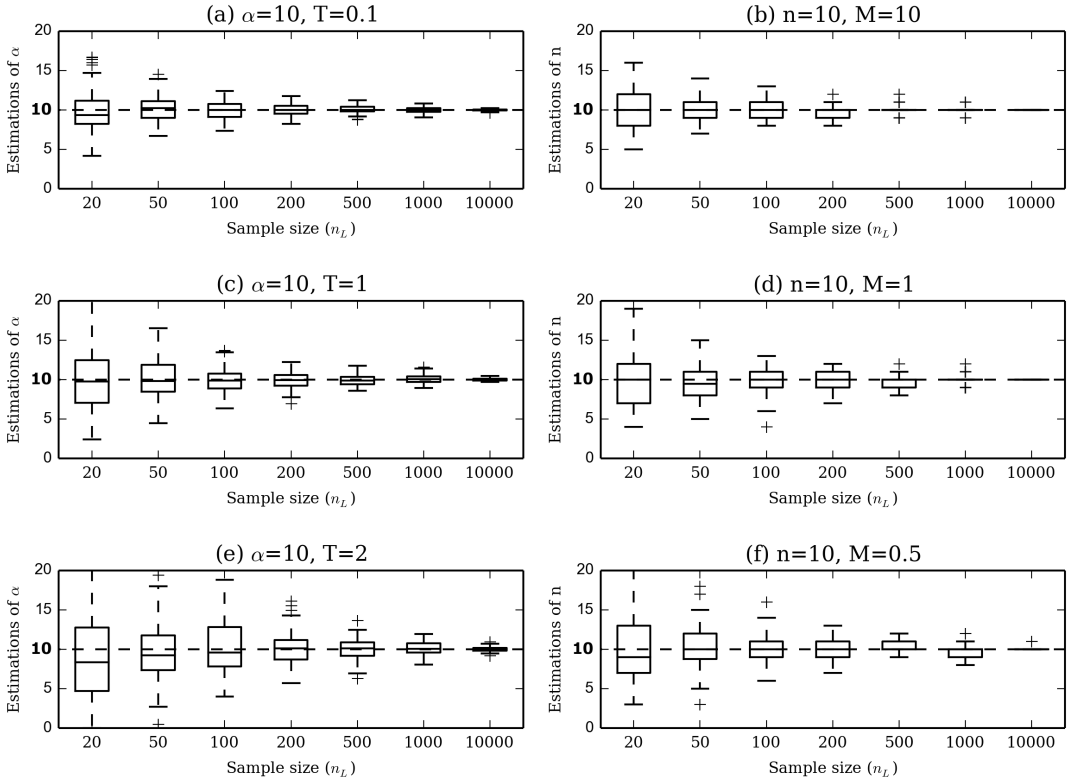


Figure 2.4: Estimation of α and n. Panels (a), (c) and (e): Estimation of α under the SSPSC model for different sample sizes and T values. Simulations performed with $\alpha = 10$ and $T = (0.1, 1, 2)$. Panels (b), (d) and (f): Estimation of n under the StSI model for different sample sizes and M values. Simulations performed with $n = 10$ and $M = (10, 1, 0.5)$.

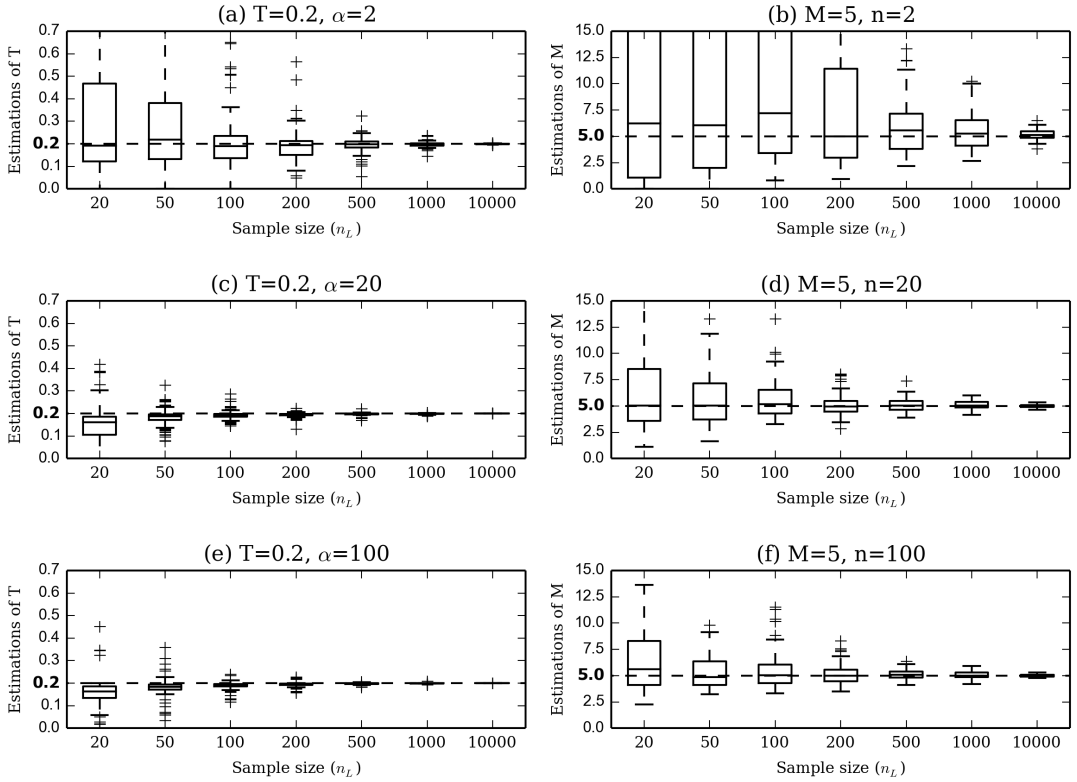


Figure 2.5: Estimation of T and M. Panels (a), (c), (e)): Estimation of T under the SSPSC model for different sample sizes and values of α. Simulations performed with $\alpha = (2, 20, 100)$ and $T = 0.2$. Panels (b), (d), (f): Estimation of M under the StSI model for different sample sizes and values of n. Simulations performed with $n = (2, 20, 100)$ and $M = 5$.

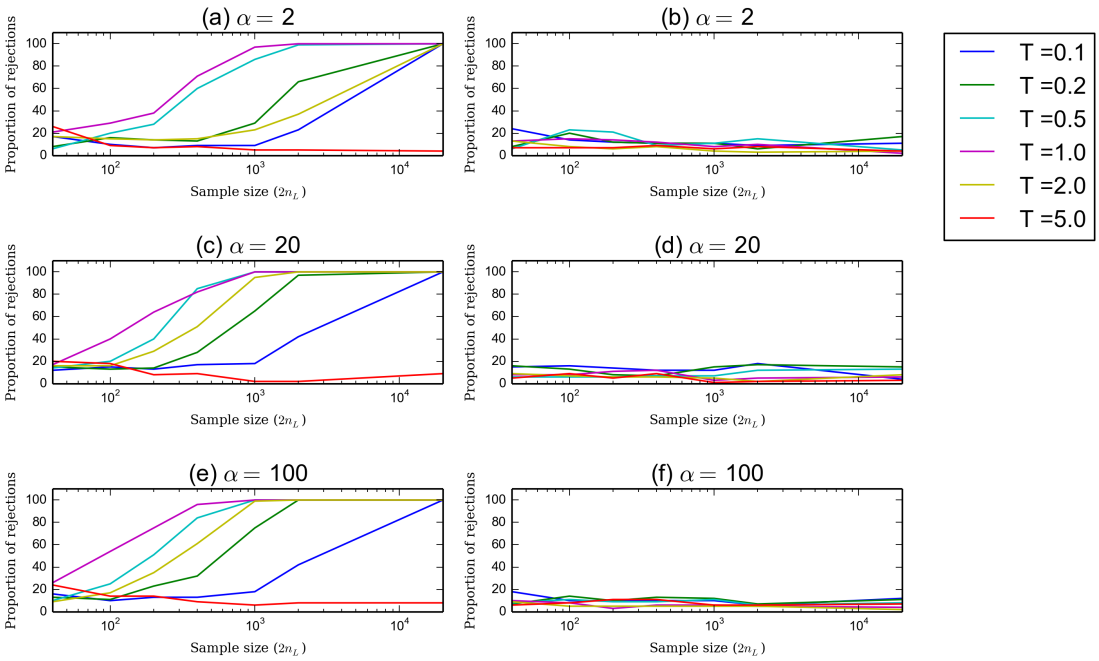


Figure 2.6: Proportion of rejected data sets simulated under the SSPSC model. Panels (a), (c) and (e): the reference model is the StSI model. Panels (b), (d), and (f): the reference model is the SSPSC, *i.e.* the model under which the data were simulated. Note that for the abscissa we used $2n_L$ instead of n_L because in order to perform the KS test it is necessary to first estimate the parameters using n_L loci and then an independent set of n_L values of T_2.

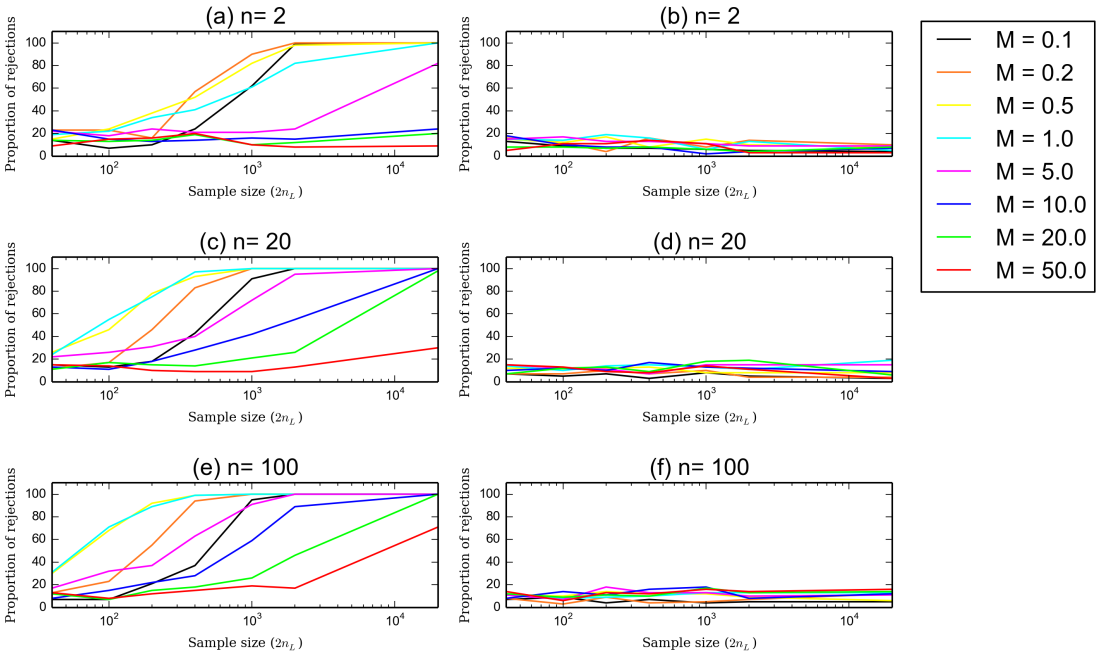


Figure 2.7: Proportion of rejected data sets simulated under the StSI model. Panels (a), (c), and (e): the reference model is the SSPSC. Panels (b), (d), and (f): the reference model is the StSI model, *i.e.* the model under which the data were simulated. Note that for the abscissa we used $2n_L$ instead of n_L because in order to perform the KS test it is necessary to first estimate the parameters using n_L loci and then an independent set of n_L values of T_2.

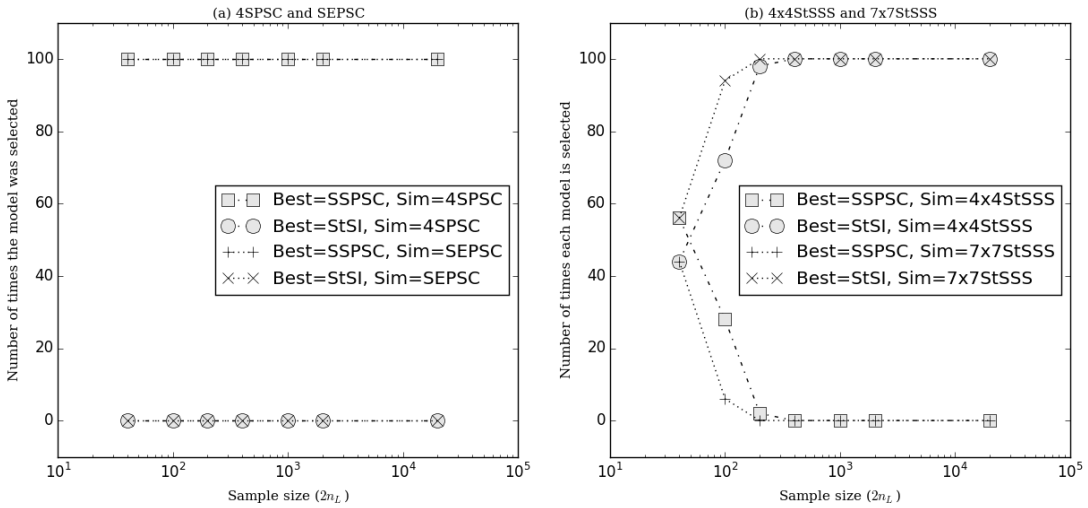


Figure 2.8: Model choice using the AIC for various demographic models. This figure shows the proportion of times the AIC selected the SSPSC (resp. the StSI) as the best model, as a function of n_L, the number of independent loci or T_2 values. For each n_L value, the experiment was repeated 100 times, and the number of times one model was chosen is plotted. In each panel we represent the model selected by the AIC as "Best" and the model under which the data were simulated as "Sim". In the left panel the data were simulated under the two models of population size change, namely the 4SPSC (4 stepwise population size changes) and the SEPSC (a single exponential population increase. In the right panel, the data were simulated under the two stepping-stone models models, the 4x4StSSS (with 4x4 islands) and the 7x7StSSS (with 7x7 islands). This figure shows that the AIC provides very good results to identify a structured model compared to a model of population size change.

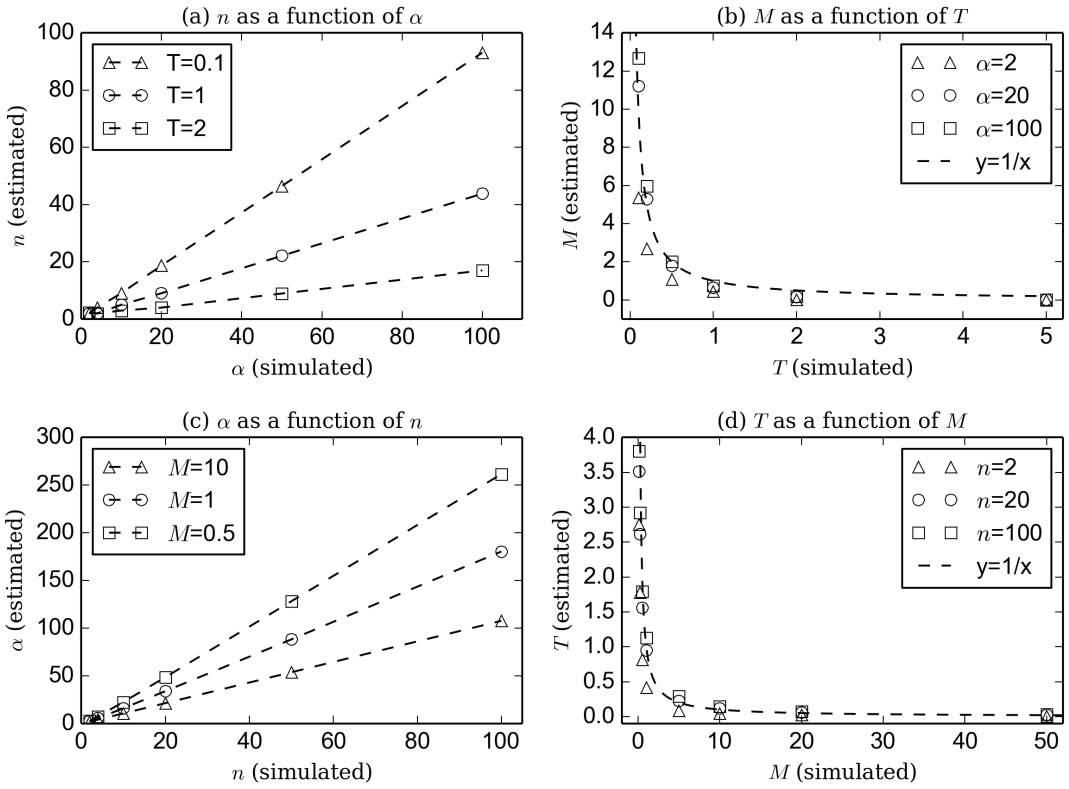


Figure 2.9: Relationships between parameters of the models

2.7 Theoretical details and perspectives

2.7.1 Derivations of distributions of T_2^{SSPSC} and T_2^{StSI}

The distribution of T_2^{SSPSC} If we denote by $\lambda(t)$ the ratio $\frac{N(t)}{N(0)}$ where t is the time scaled by the number of genes (*i.e.* units of coalescence time, corresponding to $\lfloor N(0)t \rfloor$ generations), we can compute the probability density function (*pdf*) $f_{T_2}^{SSPSC}(t)$ of the coalescence time T_2^{SSPSC} of two genes sampled in the present-day population. Indeed, the probability that two genes will coalesce at a time greater than t is

$$\mathbb{P}(T_2^{SSPSC} > t) = e^{-\int_0^t \frac{1}{\lambda(x)} dx} , \tag{2.3}$$

where

$$\lambda(x) = \mathbb{I}_{[0,T[}(x) + \alpha \mathbb{I}_{[T,+\infty[}(x),$$

and $\mathbb{I}_{[a,b[}(x)$ is the Kronecker index such that

$$\mathbb{I}_{[a,b[}(x) = \begin{cases} 1 & \text{for } x \in [a,b[\\ 0 & \text{otherwise.} \end{cases}$$

Given that the *pdf* is

$$f_{T_2}^{SSPSC}(t) = (1 - \mathbb{P}(T_2^{SSPSC} > t))'$$

Equation (1) can be rewritten as

$$\mathbb{P}(T_2^{SSPSC} > t) = e^{-t}\mathbb{I}_{[0,T[} + e^{-T-\frac{1}{\alpha}(t-T)}\mathbb{I}_{[T,+\infty[}.$$

This leads to the following *pdf*

$$f_{T_2}^{SSPSC}(t) = e^{-t}\mathbb{I}_{[0,T[}(t) + \frac{1}{\alpha}e^{-T-\frac{1}{\alpha}(t-T)}\mathbb{I}_{[T,+\infty[}(t). \tag{2.4}$$

The distribution of T_2^{StSI} An easy way to derive the distribution of the coalescence time T_2^{StSI} of two genes for our structured model, is to compute the probability that two genes are identical by descent when they are sampled from the same or from different populations. These two probabilities are respectively denoted by $p_s(\theta)$ and $p_d(\theta)$, where $\theta = 2uN$ is the scaled mutation rate, u being the *per* locus mutation rate.

Indeed, using a classical scaling argument, we can note that

$$p_s(\theta) = \mathbb{E}(e^{-\frac{\theta}{2}2T_2^{StSI}}) = \mathbb{E}(e^{-\theta T_2^{StSI}}).$$

In other words $p_s(\theta)$ is the Laplace transform of T_2^{StSI}.

We can compute this probability as follows. Taking two genes from the same island and going back in time, there are three events that may occur: a coalescence event (with rate 1), a mutation event (with rate θ) and a migration event (with rate M). Taking now two genes from different islands, they cannot coalesce and therefore only a mutation or a migration event may occur. Migration events can then bring the lineages in the same island with probability $\frac{1}{n-1}$, and in different islands with probability $\frac{n-2}{n-1}$. We thus obtain the following coupled equations:

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

and

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

By solving them, we obtain

$$p_s(\theta) = \frac{\theta+\gamma}{D} \text{ and } p_d(\theta) = \frac{\gamma}{D}$$

with

$$\gamma = \frac{M}{n-1} \text{ and } D = \theta^2 + \theta(1+n\gamma) + \gamma.$$

We can then obtain the full distribution through the Laplace transform formula, if we note that

$$p_s(\theta) = \frac{\theta+\gamma}{(\theta+\alpha)(\theta+\beta)} = \frac{a}{\theta+\alpha} + \frac{1-a}{\theta+\beta}$$

with

$$a = \frac{\gamma-\alpha}{\beta-\alpha} = \frac{1}{2} + \frac{1+(n-2)\gamma}{2\sqrt{\Delta}},$$

where

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

and

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

$\Delta = (1+n\gamma)^2 - 4\gamma$ being the discriminant of the polynomial D. Noting now that for any θ and any α we have

$$\int_0^{+\infty} e^{-\alpha s} e^{-\theta s}\, ds = \frac{1}{\theta+\alpha},$$

it is straightforward to see that the *pdf* of T_2^{StSI} is an exponential mixture:

$$f_{T_2}^{StSI}(t) = ae^{-\alpha t} + (1-a)e^{-\beta t}. \tag{2.5}$$

2.7.2 Proof of the Lemma 2.1

Lemma 2. *Given a set of n_L independent observations $\{t_1, t_2, ..., t_{n_L}\}$, the critical points of interest of the log-likelihood function $\log(\mathbb{L}_{SSPSC})$ are of the form*

$$m_a = (\alpha_a, t_a), a \in \{1, 2, ..., n_L\},$$

with

$$\alpha_a = \frac{1}{K}\left(\sum_{i=1}^{n_L} t_i \mathbb{I}_{t_a < t_i} - K t_a\right) \qquad \text{and} \qquad K = \sum_{i=1}^{n_L} \mathbb{I}_{t_i \leq t_a}.$$

Proof. Given n_L independent values $t = (t_1, t_2, ..., t_{n_L})$, the likelihood is:

$$\mathbb{L}_{SSPSC}(\alpha, T) = \prod_{i=1}^{n_L} \mathbb{L}_{t_i}(\alpha, T),$$

and taking the log:

$$\log(\mathbb{L}_{SSPSC}(\alpha, T)) = \sum_{i=1}^{n_L} \log(\mathbb{L}_{t_i}(\alpha, T)), \tag{2.6}$$

where

$$\mathbb{L}_{t_i}(\alpha, T) = \frac{1}{\alpha} e^{-T - \frac{1}{\alpha}(t_i - T)} \mathbb{I}_{T \leq t_i} + e^{-t_i} \mathbb{I}_{T > t_i}. \tag{2.7}$$

First note that:

- For $T > t_i$, $\mathbb{L}_{t_i}(\alpha, T) = e^{-t_i}$ is **constant** (with respect to (α, T)).
- If $\alpha \neq 1$ ($\alpha = 1$ means there is no change in the population's size) then $\mathbb{L}_{t_i}(\alpha, T)$ has a discontinuity at $T = t_i$.

As we are interested in the case $\alpha \neq 1$, the log-likelihood function has discontinuities at each t_i, $i = 1 \dots n_L$.

For $i \in \{0, 1, ..., n_L + 1\}$, let $C_i = \{(\alpha, T) \in \mathbb{R}^+ \times \mathbb{R}^+, t_i < T < t_{i+1}\}$ with $t_0 = 0$ and $t_{n_L+1} = +\infty$.

Now let be $C = \bigcup_{i=0}^{n_L} C_i$ (Figure 2.10). We can see that $\mathbb{L}_{SSPSC}(\alpha, T)$ is continuously differentiable in the interior of C.

Given that we do not consider negative values for α or T, we split the parameter space into two subsets C and $\mathbb{R}^2 \backslash C$. If $(\alpha, T) \in C$, taking the log and the derivative with respect to T in equation (2.7) gives:

$$\frac{\partial}{\partial T}\log\left(\mathbb{L}_{t_i}(\alpha,T)\right)=\begin{cases}-1+\frac{1}{\alpha} & \text{if } T<t_i\\ 0 & \text{otherwise.}\end{cases}$$

As we can see, if $\alpha<1$ then $\frac{\partial}{\partial T}\log(\mathbb{L}_{t_i}(\alpha,T))>0$ for all i and if $\alpha>1$ then $\frac{\partial}{\partial T}\log(\mathbb{L}_{t_i}(\alpha,T))<0$ for all i. A consequence, $\nabla\log(\mathbb{L}_{SSPSC}(\alpha,T))$ will never be zero in the interior of C if $\alpha\neq 1$. □ □

This fact suggests that the *min* and *max* values of $\mathbb{L}_{SSPSC}$ (if they exist) have the form (α,t_i).

Let's find the critical points of $\log(\mathbb{L}_{SSPSC})$ over the lines (α,t_i) with $t_i\in\{t_1,t_2,...,t_{n_L}\}$. When we fix the value of T the function becomes a function of the single variable α. If $T=t_a$ for $a\in\{1,...,n_L\}$ it follows from (2.6) that:

$$\log(\mathbb{L}_{SSPSC}(\alpha,t_a))=\sum_{i=1}^{n_L}\left[\log\left(\frac{1}{\alpha}\right)-t_a-\frac{1}{\alpha}(t_i-t_a)\right]\mathbb{I}_{t_a\leq t_i}-\sum_{i=1}^{n_L}t_i\mathbb{I}_{t_a>t_i}.$$

Denoting $K=\sum_{i=1}^{n_L}\mathbb{I}_{t_a\leq t_i}$, we then have:

$$\log(\mathbb{L}_{SSPSC}(\alpha,t_a))=K\left(\log\left(\frac{1}{\alpha}\right)-t_a\right)-\frac{1}{\alpha}\sum_{i=1}^{n_L}(t_i-t_a)\mathbb{I}_{t_a<t_i}-\sum_{i=1}^{n_L}t_i\mathbb{I}_{t_a>t_i}.$$

Let us find the zeros of the derivative in α:

$$\begin{aligned}\frac{\partial}{\partial\alpha}\log(\mathbb{L}_{SSPSC}(\alpha,t_a))=0 &\Leftrightarrow -\frac{K}{\alpha}+\frac{1}{\alpha^2}\sum_{i=1}^{n_L}(t_i-t_a)\mathbb{I}_{t_a\leq t_i}=0\\ &\Leftrightarrow -\alpha K+\sum_{i=1}^{n_L}(t_i-t_a)\mathbb{I}_{t_a\leq t_i}=0\\ &\Leftrightarrow\alpha=\frac{1}{K}\sum_{i=1}^{n_L}t_i\mathbb{I}_{t_a\leq t_i}-t_a=\alpha_a.\end{aligned}$$

Hence, the maximum value of the log-likelihood function (if it exists) is of the form:

$$m_a=\left(\frac{1}{K}\sum_{i=1}^{n_L}t_i\mathbb{I}_{t_a\leq t_i}-t_a,t_a\right),a\in\{1,2,...,n_L\}.$$

We then take $(\hat{\alpha},\hat{T})=argmax_{a\in\{1,...,n_L\}}\{\log(\mathbb{L}_{SSPSC}(m_a))\}$ as the Maximum Likelihood Estimation.

Remark: Note that the function $\mathbb{L}_{SSPSC}$ actually does not have any upper bound. Let $(t_1, t_2, ..., t_{n_L})$ be the n_L observations of T_2 sorted form the lower value to the higher value.

For $T = t_{n_L}$ we have from (2.6) and (2.7):

$$\mathbb{L}_{SSPSC}(\alpha, t_{n_L}) = \frac{1}{\alpha} e^{-\sum_{i=1}^{n_L} t_i}$$

which clearly goes to $+\infty$ as α goes to zero. Besides, let us note that $m_{n_L} = 0$! So in practice, we remove t_{n_L} from the possible values for $\tilde{T}$, our m_a remaining a good estimate for the parameters, since the probability $\mathbb{P}(t_{n_L-1} < T)$ is very low for most of the situations: less than 10^{-4} for all values of $n_L \geq 20$ if $T \leq 1$, and less than 10^{-6} for all values of $T \leq 5$ if $n_L \geq 100$.

2.7.3 Number of differences between pairs

If we assume that the scaled mutation rate is equal to θ, then the number of differences between pairs of non recombining sequences (N_d), conditioned by the value of T_2 can be computed as:

$$\mathbb{P}(N_d = k | T_2 = t) = e^{-2t\theta} \frac{(2t\theta)^k}{k!}.$$

If the density of T_2 is known, then we can compute the number of differences by taking the integral over all possible values of T_2:

$$\mathbb{P}(N_d = k) = \int_0^{+\infty} \mathbb{P}(N_d = k | T_2 = t) f_{T_2}(t) dt.$$

In the following lines we derive the distribution of N_d for the two models under study.

Number of differences in the SSPSC model Let's use the intermediate result:

$$I_k = \int_0^T t^k e^{-ct}\, dt = -\frac{1}{k} T^k e^{-cT} + \frac{k}{c} I_{k-1}$$

from which, by recursion,

$$I_k = \frac{k!}{c^k} I_0 - \sum_{i=0}^{k-1} \frac{k!}{(k-i)!} T^{k-i} \frac{e^{-cT}}{c^{i+1}},$$

with

$$I_0 = \frac{1 - e^{-cT}}{c},$$

we get:

$$\int_0^T t^k e^{-ct}\,dt = \frac{k!}{c^{k+1}} - \sum_{i=0}^{k} \frac{k!}{(k-i)!}\frac{T^{k-i}e^{-cT}}{c^{i+1}}.$$

We know that:

$$\mathbb{P}(N_d^{SSPSC} = k) = \int_0^{+\infty} \mathbb{P}(N_d = k|T_2 = t) f_{T_2^{SSPSC}}(t)\,dt,$$

which is equal to:

$$\int_0^{+\infty} e^{-2t\theta}\frac{(2t\theta)^k}{k!}\left(e^{-t}\mathbb{I}_{[0,T[}(t) + \frac{1}{\alpha}e^{-T-\frac{1}{\alpha}(t-T)}\mathbb{I}_{[T,+\infty[}(t)\right)\,dt.$$

This integral can be computed as the sum of two integrals:

$$\frac{(2\theta)^k}{k!}\int_0^T t^k e^{-(2\theta+1)t}\,dt + \int_T^{+\infty}\frac{1}{\alpha}e^{-T-\frac{1}{\alpha}(t-T)}\frac{(2\theta)^k}{k!}t^k e^{-2t\theta}\,dt.$$

The second integral can be calculated by doing:

$$\int_T^{+\infty}\frac{1}{\alpha}e^{-T-\frac{1}{\alpha}(t-T)}\frac{(2\theta)^k}{k!}t^k e^{-2t\theta}\,dt = \frac{1}{\alpha}\frac{(2\theta)^k}{k!}e^{-T(2\theta+1)}\int_0^{+\infty}(u+T)^k e^{-u(\frac{1}{\alpha}+2\theta)}\,du,$$

with

$$\int_0^{+\infty}(u+T)^k e^{-u(\frac{1}{\alpha}+2\theta)}\,du = \sum_{i=0}^{k} C_k^i T^{k-i}\int_0^{+\infty} u^i e^{-u(\frac{1}{\alpha}+2\theta)}\,du = \sum_{i=0}^{k} C_k^i T^{k-i}\frac{i!}{(\frac{1}{\alpha}+2\theta)^{i+1}}.$$

Putting all together:

$$\boxed{\mathbb{P}(N_d^{SSPSC} = k) = \frac{(2\theta)^k}{(2\theta+1)^{k+1}} + (2\theta)^k\sum_{i=0}^{k}\frac{e^{-T(2\theta+1)}T^{k-i}}{(k-i)!}\left(\frac{1}{\alpha(2\theta+\frac{1}{\alpha})^{i+1}} - \frac{1}{(2\theta+1)^{i+1}}\right)}$$

Number of differences in the StSI model Here we will use the intermediate result:

$$\int_0^{+\infty} t^k e^{-ct}\,dt = \frac{k!}{c^{k+1}}.$$

As we stated before, in the StSI case we have:

$$\mathbb{P}(N_d^{StSI} = k) = \int_0^{+\infty}\mathbb{P}(N_d = k|T_2 = t) f_{T_2^{StSI}}(t)\,dt.$$

Substituting the conditional probability and the density of T_2^{StSI}, this is equal to:

$$\int_0^{+\infty} e^{-2t\theta}\frac{(2t\theta)^k}{k!}\left(ae^{-\alpha t} + (1-a)e^{-\beta t}\right)\,dt.$$

By linearity we have:

$$a\frac{(2\theta)^k}{k!}\int_0^{+\infty} t^k e^{-(\alpha+2\theta)t}\,dt + (1-a)\frac{(2\theta)^k}{k!}\int_0^{+\infty} t^k e^{-(\beta+2\theta)t}\,dt,$$

and finally:

$$a\frac{(2\theta)^k}{k!}\frac{k!}{(\alpha+2\theta)^{k+1}} + (1-a)\frac{(2\theta)^k}{k!}\frac{k!}{(\beta+2\theta)^{k+1}}$$

Hence, the distribution of N_d^{StSI} can be written as:

$$\boxed{\mathbb{P}(N_d^{StSI}=k) = \frac{a}{\alpha+2\theta}\left(\frac{1}{1+\frac{\alpha}{2\theta}}\right)^k + \frac{1-a}{\beta+2\theta}\left(\frac{1}{1+\frac{\beta}{2\theta}}\right)^k}$$

2.7.4 Preliminary results on the number of differences

Having an explicit expression for the distribution of the number of differences (N_d) makes it possible to use a strategy analogous to the one based on T_2, in order to distinguish this two simple models based on genetic data. Thus, we can use a chi-square test to decide whether the observed data can be explained by one of these two models. It is also possible to estimate parameter of both models from real data, based on the explicit expression of the distribution function of N_d.

The development of efficient algorithms for doing parameter estimation is in progress. Also, a work is in progress in order to identify whether it is possible to distinguish the two models using a reasonable amount of genetic information, and in a reasonable amount of time. Some preliminary results indicates that the strategy presented for the distribution of T_2 could be applied to real data using the observed values of N_d.

We can see from Figure 2.11 that the method is able to distinguish both models. In order get the results shown in Figure 2.11, we did the following:

- We simulated data (vectors of n_L independent values of N_d, for $n_L \in \{40, 100, 200, 400, 1000, 200$ under a panmictic model with a bottleneck of ratio 4 occurred in different times.
- we estimated the parameters, using a MLE approach under both models, using half of the data.
- We did a chi-square test to decide whether the data (the other half of the data) is explained by the SSPSC model (or the StSI model) with the estimated parameters.

When the number of independent values of N_d is large enough ($n_L = 20000$) we can see that the chi-square test rejects the hypothesis that the data correspond to a StSI model (which is the wrong model) almost every time (Figure 2.11 left panel). However the rejection rate for the SSPSC model (which is the correct model) is low (Figure 2.11 right panel).

The accuracy in the estimation of the number of islands (n) under an n-island model is also remarkable. We can see in Figure 2.12 the estimations of n, from 20000 independent values of N_d, simulated under an n-island model with $n = 10$ and migration rate (M) equal to $0.1, 1$ and 50. For each value of M we repeated 100 times the process of simulate the data and estimate n by a Maximum Likelihood Strategy. Most of the time for $M = 0.1$ (Figure 2.12 left) the estimated value of n was the right value (10) being the minimum estimated value equal to 9 and the maximum equal to 11. For $M = 1$ and $M = 50$ (Figure 2.12 middle and right) the estimation of n was always equal to 10.

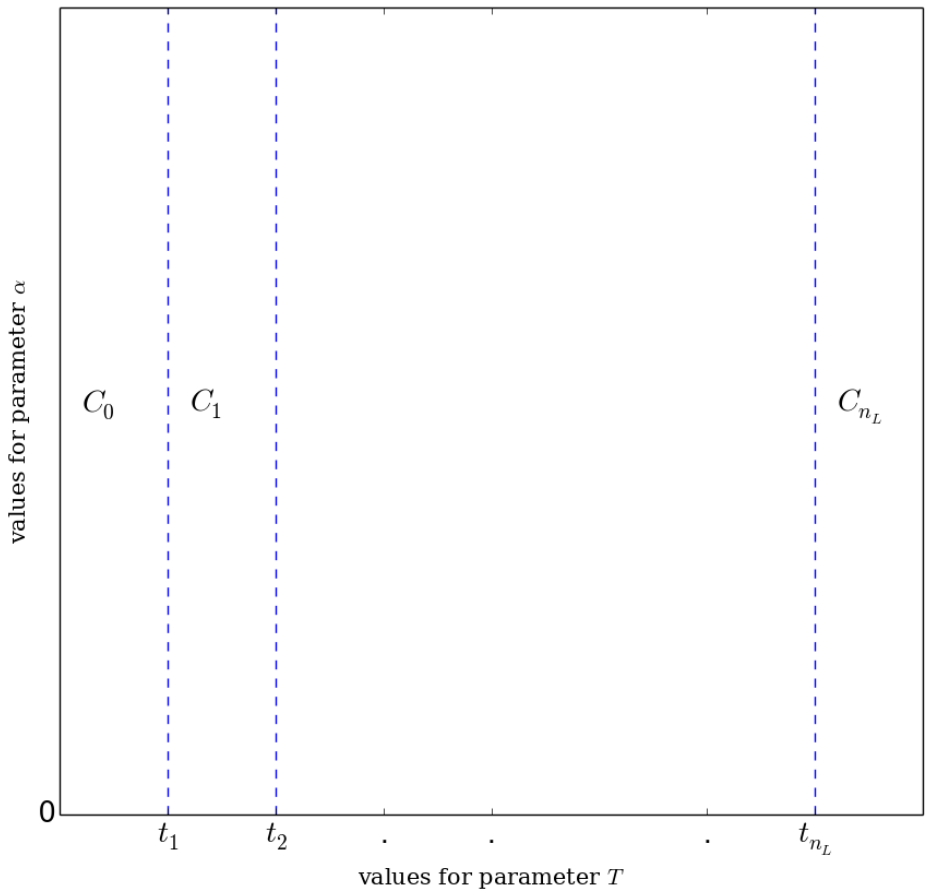


Figure 2.10: Cutting the parameter space, taking into account the discontinuities of the log-likelihood function in $\mathbb{R}^2$

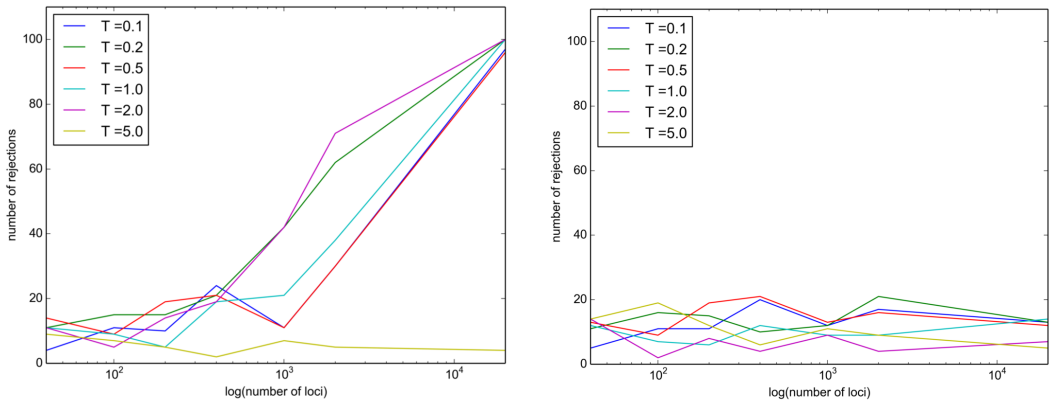


Figure 2.11: Results of a Chi-2 test using the number of pairwise differences. Data were simulated under the panmictic model with one population size change.

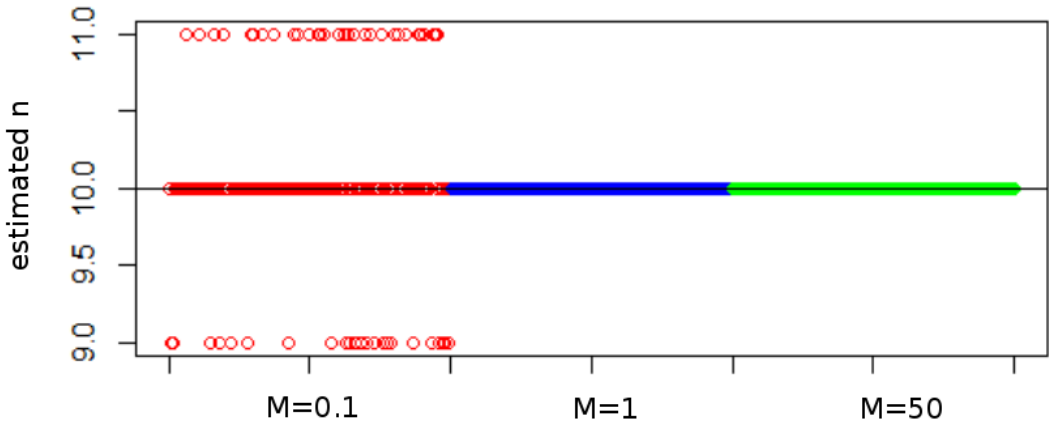


Figure 2.12: Accuracy on the estimation of n.

Chapter 3

On the importance of being structured: instantaneous coalescence rates and human evolution - Lessons for ancestral population size inference?

The hypothesis test developed in the previous chapter allows us to separate two simple models using the distribution of T_2 values. Even if the same strategy appears to have good results when applied to real data (ongoing Masters project of Alexandre Changenet), the models considered are still very simplistic. A natural question that arises is whether a similar procedure would be able to distinguish between a structured model and a panmictic model with many population size changes. In the following chapter we will see that any strategy developed to separate a structured model from a panmictic model with an arbitrary function of population size changes will fail if this strategy is based only on the distribution of T_2 values. Moreover, we will give the exact function of pseudo-population size changes (λ) which makes the distribution of T_2 under a panmictic model be identical to the one of an n-island model. This allows to predict the population size changes that will be inferred by methods assuming panmixia, when applied to a structured population. We will also note that the notion of effective size in a structured model is unclear, given that we cannot characterise this effective size by a single number, even though the population remains constant. The chapter is identical to the published work Mazet et al. (2015a).

Abstract

Most species are structured and influenced by processes that either increased or reduced gene flow between populations. However, most population genetic inference methods assume panmixia and reconstruct a history characterized by population size changes. This is potentially problematic since population structure can generate spurious signals of population size change through time. Moreover, when the model assumed for demographic inference is misspecified, genomic data will likely increase the precision of misleading if not meaningless parameters. For instance, if data were generated under an n-island model (characterized by the number of islands and migrants exchanged) inference based on a model of population size change would produce precise estimates of a bottleneck that would be meaningless. In addition, archaeological or climatic events around the bottleneck's timing might provide a reasonable but potentially misleading scenario. In a context of model uncertainty (panmixia *versus* structure) genomic data may thus not necessarily lead to improved statistical inference.

We consider two haploid genomes and develop a theory which explains why any demographic model with structure will necessarily be interpreted as a series of changes in population size by inference methods ignoring structure. We formalize a parameter, the IICR (inverse instantaneous coalescence rate), and show that it is equivalent to a population size only in panmictic models, and is mostly misleading for structured models. We argue that this issue affects all population genetics methods ignoring population structure may infer population size changes that never took place. We apply our approach to human genomic data.

3.1 Introduction

Most species are structured, and do not behave as panmictic populations (Wakeley, 1999; Harpending and Rogers, 2000; Goldstein and Chikhi, 2002; Charlesworth et al., 2003; Harding and McVean, 2004). They have been influenced by habitat fragmentation, expansion or reconnection events that either increased or reduced the amount of gene flow between local populations, as a result of climatic or anthropogenic events (Goossens et al., 2006; Quéméré et al., 2012). While genomic data offer the possibility to reconstruct with increasing precision major events in that complex history (Gutenkunst et al., 2009; Li and Durbin, 2011; Sheehan et al., 2013; Schiffels and Durbin, 2014; Liu and Fu, 2015), it is computationally very difficult to account for population structure. As a consequence, many inferential methods tend to ignore population structure (Li and Durbin, 2011; Sheehan et al., 2013; Liu and Fu, 2015). This is potentially problematic because an increasing number of studies have shown that population structure generates spurious signals of changes in population size, even when populations were stationary (Wakeley, 1999, 2001; Nielsen and Beaumont, 2009; Chikhi et al., 2010; Peter et al., 2010; Heller et al., 2013; Paz-Vinas et al., 2013; Mazet et al., 2015b). Here, we provide a simple theoretical framework which explains why any inferential method ignoring population structure will always infer population size changes as soon as populations are actually structured. In other words, this theory explains why any real demographic history, with or without structure, will necessarily and optimally be interpreted as a series of changes in population size by methods ignoring population structure.

We consider the case of two haploid genomes and we study T_2, the coalescence time for a sample of size two (*i.e.* the time to the common ancestor of two randomly sampled sequences (Herbots, 1994; Griffiths and Tavaré, 1994; Mazet et al., 2015b)). We predict the history that any coalescent-based population genetics methods ignoring structure will try to reconstruct. We introduce a parameter, which we call the IICR (inverse instantaneous coalescence rate). Since coalescence rates are expected to be inversely related to effective population sizes, it may seem natural to see the IICR as an "instantaneous population size". However, we stress that the IICR is equivalent to a population size only in panmictic models. For models incorporating population structure the IICR exhibits a temporal trajectory that can be strongly disconnected from the real demographic history (*i.e.* identifying a decrease when the population size was actually constant or increasing).

We apply our approach to simulated data and use the PSMC (Pairwise Sequentially Markovian Coalescent) method (Li and Durbin, 2011) as a reference method because it allows to reconstruct the history of a population or species from one single diploid genome. Also, this method has been applied to a wide array of

vertebrate species including reptiles (Green et al., 2014), birds (Zhan et al., 2013; Hung et al., 2014) and mammals such as primates (Prado-Martinez et al., 2013; Zhou et al., 2014), pigs (Groenen et al., 2012) and pandas (Zhao et al., 2013) and its outputs have been and typically are interpreted in terms of population size changes. However, our results are general and not specifically related to that particular method.

We then apply our approach to human data and show that an alternative model involving a minimum of three changes in migration rates can explain the PSMC results obtained by Li and Durbin (2011). The scenario that we infer represents an alternative to the population crashes and increases depicted in various population genetic studies, but is strikingly in phase with fossil data and provides a more realistic framework as several authors have suggested (Goldstein and Chikhi, 2002; Harding and McVean, 2004). Altogether we call for a major re-evaluation of what genomic data can actually tell us about the demographic history of our species. Beyond our species we argue that genomic data should be re-interpreted as a consequence of changes in levels of connection rather than simple changes in population size (see also Wakeley (1999, 2001); Harding and McVean (2004) for interesting models incorporating structure).

3.2 Models, Theory

3.2.1 Coalescence time for a sample of size 2 in a model of population size change

We consider a model of arbitrary and instantaneous population size change, where $N(t)$ represents the population size (N, in units of genes or haploid genomes) as a function of time (t) scaled by the number of genes (*i.e.* in units of coalescence time, corresponding to $\lfloor N(0)t \rfloor$ generations). We consider that $t = 0$ is the present, and positive values represent the past. Since N represents the population size in terms of haploid genomes, the number of individuals will be $N/2$ for diploid species. We can then apply the generalisation of the coalescent in populations of variable size (Griffiths and Tavaré, 1994; Donnelly and Tavaré, 1995; Tavaré, 2004). If we denote by $\lambda(t)$ the ratio $\frac{N(t)}{N(0)}$, we can then compute the probability density function (*pdf*) $f_{T_2}^{PSC}(t)$ of the coalescence time T_2 of two genes sampled in the present-day population. Indeed, the probability that two genes will coalesce at a time greater than t is

$$\mathbb{P}(T_2 > t) = e^{-\int_0^t \frac{1}{\lambda(x)}\,dx} \tag{3.1}$$

Given that

$$f_{T_2}^{PSC}(t) = (1 - \mathbb{P}(T_2 > t))' \tag{3.2}$$

we can write the *pdf* as

$$f_{T_2}^{PSC}(t) = (1 - e^{-\int_0^t \frac{1}{\lambda(x)}\,dx})' = \frac{1}{\lambda(t)} e^{-\int_0^t \frac{1}{\lambda(x)}\,dx} \tag{3.3}$$

Consequently, if we know the *pdf* of the coalescence time T_2, the corresponding population size change function $\lambda(t)$ can be computed as:

$$\lambda(t) = \frac{\mathbb{P}(T_2 > t)}{f_{T_2}^{PSC}(t)} \tag{3.4}$$

This equation may be seen as a simple rearrangement of previously known results (Griffiths and Tavaré, 1994; Tavaré, 2004), which we cited above, and to some extent it is. However, it practically means that if we only had access to a finite set of T_2 values we could in theory infer the history $\lambda(t)$ by simply computing this ratio. In the case of a model of population size change this computation is by definition giving us the actual history of population size change. We show below how this ratio can be computed for *any* demographic scenario for which T_2 distributions can be derived or simulated. And it is this computation for other models that significantly changes the outlook to genetic data and coalescence rates.

3.2.2 Instantaneous coalescence rate for a sample of size 2

If we consider now the coalescence time of two genes sampled in a population under an arbitrary model, whichever model this may be (structured or not, with population size change or not, etc.), and if we assume that we know its *pdf*, $f_{T_2}(t)$, it is straightforward to compute the ratio $\lambda(t)$ of equation (3.4)

$$\lambda(t) = \frac{\mathbb{P}(T_2 > t)}{f_{T_2}(t)} \tag{3.5}$$

Let us now denote $g(t) = \mathbb{P}(T_2 > t)$. We then have by definition $f_{T_2}(t) = -g'(t)$, hence

$$\frac{1}{\lambda(t)} = -\frac{g'(t)}{g(t)} = -\log(g(t))' \tag{3.6}$$

from where we get, since $g(0) = 1$,

$$g(t) = e^{\log(g(t))} = e^{-\int_0^t \frac{1}{\lambda(x)}\,dx} \tag{3.7}$$

It therefore follows that the *pdf* $f_{T_2}(t) = -g'(t)$ can always be written as

$$f_{T_2}(t) = \frac{1}{\lambda(t)} e^{-\int_0^t \frac{1}{\lambda(x)}\,dx} \tag{3.8}$$

even if the so-computed function $\lambda(t)$ has nothing to do with any population size change.

In other words, for any given model, there always exists a function $\lambda(t)$ which *explains* the coalescence time distribution of this model for a sample of size two, $f_{T_2}(t)$. The *pdf* of T_2 can thus always be written as a function of $\lambda(t)$ as in equation (3.8), exactly as if the model under which the data were produced was *only defined by population size changes*. This function $\lambda(t)$ is a fictitious or spurious population size change function whose coalescence time T_2 would *mimic* perfectly the demographic model.

Now, if we define $\mu(t)$ as

$$\mu(t) = \frac{1}{\lambda(t)} = \frac{f_{T_2}(t)}{\mathbb{P}(T_2 > t)} \tag{3.9}$$

it should be natural to see $\mu(t)$ as an *instantaneous coalescence rate*, as it represents the probability that two lineages which have not yet coalesced at time t (as expressed by the denominator), will do so in an infinitesimal amount of time starting at t (as expressed in the numerator). Another way to realize it is to use theoretical results and terminology from reliability theory. If we note that T_2 can be seen as a *lifetime*, then, we can also note that the quantity $\mu(t) = \frac{1}{\lambda(t)}$, known as the *hazard function* or *failure rate* in the reliability engineering community, represents the instantaneous rate of failure of a system at time t (see for instance Ruegg (1989) or Klein and Moeschberger (2003)). The term *instantaneous* is central and we show in the next section that it is crucial for the interpretation of structured models.

3.2.3 Linking population structure and population size change

We now consider a model of population structure such as the classical symmetric *n-island* model (Wright, 1931), where we have a set of n islands (or demes) of constant size N, interconnected by gene flow with a migration rate m, where $\frac{M}{2} = Nm$ is the number of immigrants (genes) in each island every generation. The total number of genes or haploid genomes in the whole metapopulation is nN and it is therefore constant. Again, N is the number of haploid genomes, and $N/2$ the number of diploid individuals.

Under this model we can write the *pdf* for T_2 (see Herbots (1994); Wilkinson-Herbots (1998); Mazet et al. (2015b) for details and Bahlo and Griffiths (2001) for related results and Charlesworth et al. (2003) for an insightful review) by

considering the cases when the two genes are sampled from the same (s) or from different (d) demes.

$$f_{T_2^s}^{StSI}(t) = ae^{-\alpha t} + (1-a)e^{-\beta t} \tag{3.10}$$

$$f_{T_2^d}^{StSI}(t) = ce^{-\alpha t} - ce^{-\beta t} \tag{3.11}$$

where

$$a = \frac{\gamma - \alpha}{\beta - \alpha}, \quad c = \frac{\gamma}{\beta - \alpha} \tag{3.12}$$

and where $-\alpha$ and $-\beta$ are the roots of the polynomial

$$\theta^2 + \theta(1 + n\gamma) + \gamma \tag{3.13}$$

whose discriminant is $\Delta = (1 + n\gamma)^2 - 4\gamma$, and therefore

$$\alpha = \frac{1}{2}\left(1 + n\gamma + \sqrt{\Delta}\right) \tag{3.14}$$

and

$$\beta = \frac{1}{2}\left(1 + n\gamma - \sqrt{\Delta}\right) \tag{3.15}$$

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

Now let us consider a hypothetical demographic history characterized by population size changes but without any population structure. For that history to explain the data generated by a model of population structure, this hypothetical demographic history will correspond to the function $\lambda(t)$ as defined by equation 3.5. Thus, in the case of two haploid genomes sampled in the same deme (a most reasonable assumption for a diploid individual) we get:

$$\lambda_s(t) = \frac{\mathbb{P}(T_2 > t)}{f_{T_2^s}^{StSI}(t)} = \frac{\frac{a}{\alpha}e^{-\alpha t} + \frac{1-a}{\beta}e^{-\beta t}}{ae^{-\alpha t} + (1-a)e^{-\beta t}} = \frac{(1-\beta)e^{-\alpha t} + (\alpha - 1)e^{-\beta t}}{(\alpha - \gamma)e^{-\alpha t} + (\gamma - \beta)e^{-\beta t}} \tag{3.16}$$

It is then trivial to compute the function $\lambda_s(t)$ for any set of parameters n and M. Figure 3.1 shows for instance in panel (a) the corresponding curves for $n = 50$ and M values between 0.1 and 50. As expected (Chikhi et al., 2010; Mazet et al., 2015b) we observe a (fictitious) population decrease from a large hypothetical ancestral population of size N_a^h to a smaller hypothetical current population of size N_c^h. Note that $\lambda_s(t)$ is a population size ratio, which does not provide absolute values of the effective population size. In our case, it is

however trivial to show that for t sufficiently close to 0, we find that $\lambda_s(t) = 1$ and hence it follows that $N_c^h = N$, the size of a deme. Indeed, at the time of sampling, the coalescence history for two genes sampled from the same deme is mostly dependent on the size of the local deme. Interestingly, this is true for any value of M. Figure 3.1 indicates that as M becomes larger, $N_a^h = N \lim_{t\to+\infty} \lambda_s(t)$ becomes closer to nN, represented by the horizontal dashed line. This is expected: when the migration rate increases the whole set of populations behaves less and less like a structured model and increasingly like a single random mating population of size nN. Several authors have shown that under the strong migration condition, it is possible to define a coalescent effective population size towards which the structured population tends (Sjödin et al., 2005; Wakeley and Sargsyan, 2009). Panel (b) shows indeed that when M is very high ($M = 100$ and $M = 500$) the n-island model behaves as a population characterized by a constant size until the very recent past. For instance, when $M = 500$, $\lambda_s(t)$ only drops at time $t = 0.02$, which for $N = 100$ would correspond to 2 generations ago. In other words, the strong migration assumption implicitly assumes that the bottleneck seen in our results is so recent that it can be neglected. Using the terminology introduced by Wakeley (1999), it assumes that the scattering phase is very short. Altogether our results provide a more general framework which allows us to easily incorporate the strong migration assumptions.

Coming back to panel (a) we also note that as M decreases, the fictitious bottleneck becomes older and the ancestral population becomes larger, for a constant value of n, the number of islands. We can derive the asymptotic coalescent effective size of this n-island model by computing the limit of $\lambda(t)$ when t goes to infinity, and find that, since $0 < \beta < \alpha$,

$$N_a^h = N \lim_{t\to+\infty} \lambda_s(t) = N\frac{\alpha - 1}{\gamma - \beta} = \frac{N}{\beta}, \tag{3.17}$$

where we recall that β was the smallest of the roots found above (equation 3.15). By developing equation 3.15, we find

$$\beta = \frac{1}{2}\left(1 + \frac{n}{n-1}M - \sqrt{\left(1 + \frac{n}{n-1}M\right)^2 - \frac{4M}{n-1}}\right) \tag{3.18}$$

Here we can see that for large values of M, $\lambda_s(t)$ is close to

$$N_a^h = N(n + \frac{(n-1)^2}{nM}) \tag{3.19}$$

This is the nucleotide diversity effective size computed in Nei and Takahata (1993) for the n-island model.

If we now perform the same analyses and computations for the case where the haploid genomes are sampled from different demes leads to the following result:

$$\lambda_d(t) = \frac{\frac{1}{\alpha}e^{-\alpha t} - \frac{1}{\beta}e^{-\beta t}}{e^{-\alpha t} - e^{-\beta t}} = \frac{\beta e^{-\alpha t} - \alpha e^{-\beta t}}{\gamma e^{-\alpha t} - \gamma e^{-\beta t}} \tag{3.20}$$

Here the population dynamics is inverted, and we observe a fictitious population expansion. Figure 3.2.3 shows some plots of $\lambda_d(t)$ for different values of M. This is in agreement with several previous studies which noted that when sampling is carried out across demes the bottleneck signal either disappears or can be replaced by a population expansion signal (Peter et al., 2010; Chikhi et al., 2010; Heller et al., 2013). We note that $\lim_{t\to 0}(\lambda_d(t)) = +\infty$. The two lineages being in different demes at time $t = 0$, it is by definition impossible for them to coalesce in the very recent past, since a migration event has first to occur. Let us note also that $\lim_{t\to\infty}(\lambda_d(t)) = \frac{1}{\beta}$ as for λ_s.

Our results, as expressed by equations (3.16) and (3.20), stress the difficulty in defining an effective size for a structured population, because a structured population has properties that a non structured population does not have. It behaves like a non-structured population that changes in size. The IICR is therefore what connects the two (structured and panmictic) models. As a consequence, there is no overwhelming reason to summarize its properties by one single number when it actually is defined either by a number of islands and a migration rate, or by a full trajectory of effective sizes. We point towards the studies of Sjödin et al. (2005) and Wakeley and Sargsyan (2009) for models and conditions under which an effective size can be defined. What we wish to stress is that the theory presented here provides a general framework for explaining and predicting population size changes that population genetics methods will infer. Below, we illustrate how this can be applied to simple and complex structured models and we also predict the population size changes that methods ignoring structure will infer. Given that $\lambda(t)$ does not necessarily correspond to actual changes in N_e we introduce the inverse instantaneous coalescence rate or IICR, which we will use for the rest of the manuscript instead of $\lambda(t)$. The reason for this is that the IICR is only equivalent to an instantaneous coalescent N_e in the case of models without structure. For other models, it is, in the absence of a better term, the inverse of an instantaneous coalescence rate. The IICR is of course by definition a function of time and implicitly leads us to consider a trajectory rather than a single value even for constant size models such as the n-island model.

3.2.4 Application to simulated and real data

In order to illustrate how an observed distribution of T_2 values can be used to infer the IICR we carried out simulations under *structured* and *unstructured* scenarios. Data were simulated using the *ms* software (Hudson, 2002). For each scenario, we simulated independent values of T_2 and used them to estimate the IICR at various time points t_i, as follows:

$$\widehat{IICR(t_i)} = \frac{1 - \widehat{F_{T_2}(t_i)}}{\widehat{f_{T_2}(t_i)}} \tag{3.21}$$

where $\widehat{F_{T_2}(t_i)}$ is the estimated or empirical cumulative distribution function of T_2 and $\widehat{f_{T_2}(t_i)}$ is an estimated approximation of its density around t_i. The two scenarios of population size change without structure were simulated with the following *ms* commands: *ms 2 100 -T -L -G -16.094 -eG 0.1 0.0* for the exponential population size change (Figure 3.3, panel (a)) and *ms 2 100 -T -L -eN 0.01 0.1 -eN 0.06 1 -eN 0.2 0.5 -eN 1 1 -eN 2 2* (Figure 3.3, panel (b)) for the stepwise population size change.

In addition, for the scenarios involving population structure (Figures 3.4 and 3.5) we simulated both T_2 values and DNA sequences assuming an n-island model with $n = 10$ demes of size of $N = 1000$ haploid genomes each (*i.e.* 500 diploids), and a mutation rate of $\mu = 10^{-8}$. We then computed the empirical IICR from the T_2 values, and did a PSMC analysis using the corresponding DNA sequences. The *ms* commands used to produce the data for a model with three changes in migration rates was *ms 2 100 -t 600 -r 120 30000000 -I 10 2 0 0 0 0 0 0 0 0 0 1 -eM 3 5 -eM 6 0.8 -eM 15 5 -p 8* and *ms 2 100 -t 600 -r 120 30000000 -I 10 2 0 0 0 0 0 0 0 0 0 0 1 -eN 1 0.5 -p 8* for a model in which deme sizes doubled (and hence the metapopulation too). We also simulated scenarios with a 10- and a 50-fold deme size increase. We either kept M, the number of migrants, or m, the migration rate, constant after the changes in N (supplementary figures). In addition we simulated a scenario where the deme size varied according to a complex step function, and inferred the IICR under various migration rates (see supplementary figures).

For the comparisons with the analyses of the human data we assumed the mutation rate used by Li and Durbin (2011), namely $\mu = 2.5 \times 10^{-8}$. These authors note that the PSMC is not expected to give reliable estimates of recent population sizes (*i.e* less than 10 KY in humans), and we therefore carried out simulations with and without a recent demographic expansion following the Neolithic transition. The simulations incorporating a recent increase in deme size in humans produce PSMC and IICR profiles similar to the red line whereas the lack of a recent increase produce a curve that is flat in the recent past (see supplementary figures). For simplicity, the genomic data for the scenario with three migration rate changes were

simulated assuming $n = 10$ demes. The *ms* command used was *ms 2 100 -t 1590 -r 318 30000000 -I 10 2 0 0 0 0 0 0 0 0 0 0 0.55 -eM 4.5 4 -eM 18.0 0.55 -eM 47.5 0.85*. This command simulates an *n-island* model $n = 10$ islands, of size $N = 1060$ haploid genomes or 530 diploids. A generation time of 25 years and a mutation rate $\mu = 2.5 \times 10^{-8}$ were assumed as in Li and Durbin (2011). Following these authors we simulated 100 independent 30 MB long "chromosomes" which were then used together to represent the full 3 GB long human genome. Under that scenario, the scaled mutation is $\theta = 4 \times 530 \times 2.5 \times 10^{-8} \times 30 \times 10^6 = 1590$. Given that each island has 530 diploid individuals, the metapopulation is composed by 5300 diploid individuals. In *ms* commands, the migration rate and time are scaled in units of the diploid deme size. The number of migrants exchanged was $M = 0.55$ in the recent past and $M = 0.85$ in the most ancient past, and changed at various times indicated by the eM flag in the *ms* command. Going from the past to the present, the *ms* commands thus simulates the following demographic events: M decreased from 0.85 to 0.55 around $47.5 \times 4 \times 530 \times 25 = 2,517,500$ years ago, then M increased from 0.55 to 4.00 approximately $18 \times 4 \times 530 \times 25 = 954,000$ years ago, and finally M decreased $4.5 \times 4 \times 530 \times 25 = 238,500$ years ago from 4.00 to 0.55. After that M remained constant. Moreover, in addition to scenarios where the deme size never changed we also simulated scenarios with a rapid increase in deme size $0.25 \times 4 \times 530 \times 25 = 13,250$ years ago by a factor 40, to represent the Neolithic transition. The figures without this change are in the supplementary material.

3.3 Results

3.3.1 Predicting the inferred demographic history of non structured and structured populations: illustrations by simulations

Figure 3.3 shows the results for non structured populations that were subjected to various histories of population size change. The left-hand panel shows a population that experienced an exponential decrease from a previously constant size ancestral population. As expected, the blue solid line obtained using the full theoretical T_2 distribution is identical to the simulated history of population size changes (*i.e.* the *real* population size changes). The stepwise red solid line represents the empirical IICR. The number of t_i values or steps can be changed depending on the precision that one wishes to reach and the total number of T_2 values. We chose values similar to those typically used in recent genomic studies for comparison (Zhao et al., 2013; Zhan et al., 2013; Zhou et al., 2014) but a much greater precision can be achieved under our framework. The right-hand panel shows similar results but

for a population that went through various stepwise population size changes. This shows the remarkable match between the theoretical and empirical IICR curves and the simulated history. When a population is not structured the IICR will exactly match the real history in terms of population size changes.

Figure 3.4 is similar to Figure 3.3 but with structured populations: we sampled two haploid genomes under the n-island model, with $n = 10$ and $M = 1$. Panel (a) shows the results when the genomes were sampled in the same deme (a single diploid individual) whereas panel (b) shows the results when the two haploid genomes were sampled in different demes. These figures show again that the empirical and theoretical IICR distributions match each other. Moreover they predict the population size change history inferred by the PSMC. This suggests that the PSMC does not infer a population size change but the IICR and estimates it rather well. Finally, the IICR and the PSMC identify a (spurious) population decrease or increase depending on the sampling scheme even though the total number of haploid genomes was constant (horizontal dashed line representing the *real population size*). These results are in agreement with several studies showing that different sampling strategies applied to the same set of populations may lead to infer quite distinct demographic histories (Chikhi et al., 2010; Heller et al., 2013) even though they used different methods. Whereas the effect described by Heller et al. (2013) was observed using the Bayesian Skyline Plot method (Drummond et al., 2005), Chikhi et al. (2010) used the msvar approach of Beaumont (1999).

While Figures 3.3 and 3.4 illustrate and validate the theory developed in previous sections using two models (the n-island and population size change) for which the T_2 distribution is known, our approach to estimate the IICR is still valid when we have values of T_2 but the distribution is not known. This can happen for models that can be simulated but for which no analytical results exist (Figure 3.5). In panel (a) of Figure 3.5, we considered an n-island model with $n = 10$ demes where the total population size remained constant (each deme had a size of $N = 1000$ haploid genomes or $N/2 = 500$ diploids) but migration rates changed at three different moments in the last $30,000$ generations, as indicated by the vertical arrows. This scenario mimics a set of populations whose connectivity is changing due to fragmentation or reconnection of habitat either due to climatic or anthropogenic effects (Goossens et al., 2006; Quéméré et al., 2012). The demographic history reconstructed by the PSMC matches again the history predicted by the empirical IICR, but it is strikingly different from the actual size of the metapopulation (horizontal line). Whereas the total population size was constant throughout, the reconstructed history suggests that the population expanded and contracted on at least two occasions. A more serious issue arises from the fact that the population size changes inferred by the PSMC do not appear to match the times at which the migration rates changed, at least at the level of precision provided by the PSMC.

For instance, the last change in migration rate, M1, occurred $6,000$ generations in the past. Instead, the PSMC infers a population expansion and contraction after that event. Panel (b) corresponds to a scenario in which the size of all demes doubled $2,000$ generations before the present. Here the striking result comes from the fact that whereas the population size doubled (black broken line) the IICR and PSMC would suggest a continuous population decrease over a very long period, whose timing has again little to do with the actual history of the population. The population size change is thus missed by the PSMC. See Supplementary figures for cases where the population increased by a factor 10 and 50 and where either M or m was constant. Altogether this figure and the associated supplementary figures suggest that changes in migration patterns or changes in deme size may be misinterpreted by population genetics methods that ignore population structure, and that there is a need for methods able to identify population structure from population size change (see Peter et al. (2010); Chikhi et al. (2010); Heller et al. (2013); Mazet et al. (2015b)).

3.3.2 A tentative re-interpretation of human past demography: on the importance of being structured

In their study Li and Durbin (2011) applied the PSMC to genomic data obtained from humans and inferred a history of population size changes. As demonstrated above, what the PSMC estimates is the IICR which does not necessarily correspond to real population size changes, but may also arise from a model with changes in migration rates. To illustrate this we applied our approach to identify an island model with constant population size reproducing closely the IICR obtained by Li and Durbin (2011). For simplicity we arbitrarily assumed that the number of islands was $n = 10$, and that there were three changes in migration rates as this is the minimum number of changes required to obtain an IICR curve with two humps, assuming a constant deme size. We propose a history in which migration rates (M_i, $i = 1, 2, 3, 4$) changed at three moments (T^i, $i = 1, 2, 3$), and where M_1 corresponds to the number of migrants exchanged between demes each generation during the period between the present and T^1. More specifically, we found a change in migration rates (from $M_4 = 0.85$ to $M_3 = 0.55$) around $T^3 = 2.52$ million years (MY) ago, then a major increase (from $M_3 = 0.55$ to $M_2 = 4$) around $T^2 = 0.9 - 1.0$ MY and finally a major decrease (from $M_2 = 4$ to $M_1 = 0.55$) around $T^1 = 0.23 - 0.25$ MY ago. In other words our results would suggest changes in connectivity at the start of the Lower Pleistocene (dated at 2.58 MY), which corresponds to the emergence of the genus *Homo*. The most striking change corresponds to major increase in connectivity just before the transition between the Lower and Middle Pleistocene (dated at 0.78 MY). We find that the

Middle Pleistocene is characterized by high and sustained gene flow. Finally, connectivity abruptly decreases at $210 - 230$ KY ago just before the earliest remains of anatomically modern humans *Homo sapiens* at *ca.* 200 KY.

3.4 Discussion

3.4.1 The IICR and the PSMC

In this study we have shown that it is always possible to find a demographic history involving *only* population size changes that perfectly explains any distribution of coalescence times T_2, even when this distribution was actually generated by a model in which there was no population size change. To illustrate this we first focused on a simple n-island model for which the *pdf* of T_2 can be derived, and obtained an analytic formula of the fictitious population size change history, named IICR (inverse instantaneous coalescence rate), as a function of the number of islands and the migration rate of the model. We also showed that the IICR can be computed for any (neutral) model from any observed distribution of T_2 values. We showed that the empirical and theoretical IICRs were identical when the latter could be obtained. We then obtained the empirical IICR under models involving changes in migration rates or in deme size. This suggests that, at least for a sample of size 2, even an infinite amount of genetic data from independent loci alone may not allow to distinguish structure and population size change models. Also, the history of population size changes in Figure 3.5 would suggest that four demographic changes occured, two expansions and two contractions, whereas only three changes of the migration rate were actually simulated.

The theory presented here is simple and general. It allows us to predict the IICR and state that any method ignoring population structure will try to estimate the IICR. In the case of complex demographic histories with population structure, interpreting the IICR as a population size or a ratio of population sizes can be misleading. To clarify the difference between the IICR and an effective population size we can consider the following rationale. If a structured population could be summarized by a single N_e then a change in gene flow should be matched by a simultaneous change in N_e. In that case, changes in N_e would be misleading (since the size would not change) but their timing might still be meaningful. For instance a "hump" inferred using diCal or the PSMC could be easily translated into a change in gene flow patterns. In such a case, we could re-interpret the changes in N_e by saying, for each hump, that gene flow decreased and then increased again. What the IICR shows is that it is not that simple. The fact that a structured model can only be summarized by a *trajectory* of spurious population sizes means that the timing of changes in migration rates will interact in a complex manner hence

generating IICR profiles that may be only loosely related with population-related events. This can be seen in Figures 3.5 and 3.6 (and the supplementary figures).

These results do not invalidate the use of panmictic models for the reconstruction of population history as long as population structure can indeed be neglected (supplementary figures), but it certainly stresses the need for caution in the interpretation of this history. When Li and Durbin published their landmark study in 2011 they showed for the first time that it was possible to reconstruct the demographic history of a population by using the genome of a single diploid individual (Li and Durbin, 2011). It was a remarkable feat based on the SMC model introduced by McVean and Cardin (2005b). Its application to various species (Prado-Martinez et al., 2013; Zhou et al., 2014; Groenen et al., 2012; Zhao et al., 2013; Green et al., 2014; Zhan et al., 2013; Hung et al., 2014) has been revolutionary and led to the development of new methods (Sheehan et al., 2013; Schiffels and Durbin, 2014; Liu and Fu, 2015). However, the increasing number of studies pointing at the effect of population structure (Leblois et al., 2006; Nielsen and Beaumont, 2009; Chikhi et al., 2010; Heller et al., 2013; Paz-Vinas et al., 2013) or changes in population structure (Wakeley, 1999, 2001; Wakeley and Aliacar, 2001; Städler et al., 2009; Broquet et al., 2010; Heller et al., 2013; Paz-Vinas et al., 2013) in generating spurious changes in inferred population size suggested that new models should be analysed that can incorporate population structure (Goldstein and Chikhi, 2002; Harding and McVean, 2004). For instance, Mazet et al. (2015b) have recently shown that genomic data from a single diploid individual can be used to distinguish an n-island model from a model with a single population size change. Their likelihood-based approach uses the distribution of coalescence times for a sample of size two (T_2). This study represents an interesting alternative since it should be possible to determine whether a model of population structure is more likely than a model of population size change to explain a particular data set. The approach of Mazet et al. (2015b) is however limited to a very simple model of population size change. Demographic models inferred by several recent methods (Li and Durbin, 2011; Schiffels and Durbin, 2014; Sheehan et al., 2013; Liu and Fu, 2015) are not limited to one population size change. They are thus more realistic, and, as we have shown here this comes at a certain price. Since they allow for several tens of population size changes, they mimic more precisely the genomic patterns arising from structured models. Therefore, they reconstruct a demographic history that can optimally explain any particular pattern of genomic variation only in terms of population size changes. As we have shown here, and until we can separate models (see below) this casts doubts on any history reconstructed from genomic data by the above-mentioned approaches. Indeed, if any pattern of (neutral) genomic variation can be interpreted efficiently in terms of population size changes, then how can we identify the cases where the observed genomic data were not generated by

population size changes?

Li and Durbin (2011) acknowledged that one should be cautious when interpreting the changes inferred by their method. For instance, they showed (see their Supplementary Materials) that when one population of constant size N splits in two half sized populations that later merge again, their method will identify a change of N even though N actually never changed. Still, their method is implicitly or explicitly used and interpreted in terms of population size changes, including by themselves. There are therefore several issues that need to be addressed. One issue is to determine whether it is possible to separate models of population size change from models of population structure (Mazet et al. (2015b), see below). When population structure can be ignored, our results actually contribute to the validation of the PSMC (supplemental figures). We found that the PSMC performed impressively well and generally reconstructed the IICR with great precision. It is therefore at this stage one of the best methods (Sheehan et al., 2013; Schiffels and Durbin, 2014; Liu and Fu, 2015) published so far and remains a landmark in population genetics inference.

3.4.2 The IICR: towards a critical interpretation of effective population sizes

The concept of effective size is central to population genetics. It allows population geneticists to replace complex real-world populations by equivalent and simpler Wright-Fisher populations *that would have the same "rate of genetic drift."* (Wakeley and Sargsyan, 2009). The concept is however far from trivial and it is not always clear what authors mean when they mention the N_e of a particular species or population, as rightly noted by Sjödin et al. (2005) among others. Several N_es have been defined depending on the property of interest (inbreeding, variance in allele frequency over time, etc.) and its relationship to genetic drift (Wakeley and Sargsyan, 2009). This is a complex issue which we do not aim at reviewing or discussing in detail here.

The IICR is related to the coalescent N_e (Sjödin et al., 2005; Wakeley and Sargsyan, 2009) but it is explicitly variable with time. Given that most species are likely to be spatially structured, interpreting the IICR as a simple (coalescent) effective size may generate serious misinterpretations.

The IICR is a trajectory of instantaneous "population sizes" which fully explains complex models without loss of information. The circumstances under which this trajectory can indeed be appropriately summarized by one effective population size is still to be determined and will depend on the questions asked and the amount of markers used. For instance, for "strong migration scenarios" ($M = 500$ and $M = 100$) the inferred population size changes are recent and abrupt, and

the period during which the population was stationary will be significant in generating patterns of genetic diversity (Wakeley, 1999, 2001; Wakeley and Aliacar, 2001; Charlesworth et al., 2003; Wakeley and Sargsyan, 2009). However, even for such cases of low genetic differentiation ($F_{ST} \approx 1/2001 = 0.0005$ and $F_{ST} \approx 1/401 = 0.0025$, respectively), the spurious population size drop could perhaps be detected with genomic information. For $M = 100$ the population size decrease starts between $t = 0.05$ and $t = 0.10$, which for $N = 100$ to $N = 1000$ could correspond to values between 5 to 100 generations ago, respectively. In other words, an n-island model may actually behave differently from a WF model even under some "strong migration" conditions. The approximation will therefore be valid for some questions and data sets, and invalid for others (Charlesworth et al., 2003; Wakeley and Sargsyan, 2009). Note also that for very low migration rates ($M = 0.1$, $M = 0.2$, corresponding to very high $F_{ST} \approx 0.71$ and $F_{ST} \approx 0.56$, respectively) the recent history is also characterized by a stationary IICR. Most genes will then coalesce within demes and only a small proportion will provide information on the ancient IICR values and therefore on population structure (see Mazet et al. (2015b)).

3.4.3 The IICR and the complex history of species: towards a critical re-evaluation of population genetics inference

The PSMC has now been applied to many species, generating curves that are very similar to those represented in Figure 3.5. In panel (a) the population size changes detected by the PSMC were not correlated in a simple manner to the changes in gene flow or deme size. This is likely the result of two factors. First, a structured population cannot always be summarized by a single number. Second, the PSMC requires a discretized distribution of time which may lead to missing abrupt changes such as those simulated here. For real data sets where changes in migration rates or in population size may be smoother, this may not be so problematic. For the human data, assuming a simple model of population structure we inferred periods of change in gene flow which correspond to major transitions in the recent human evolutionary history, including the emergence of anatomically modern humans. Given that humans are likely to have been subjected to a complex history of spatial expansions and contractions and changes in the levels of gene flow (Wakeley, 1999; Harpending and Rogers, 2000; Wakeley, 2001; Goldstein and Chikhi, 2002; Harding and McVean, 2004), our results are necessarily simplistic but suggest that a re-interpretation of panmictic models may be needed and possible. Our results are at odds with a history of population crashes and increases depicted in various population genetic studies, but it is in phase with fossil data and provides

a more realistic interpretation framework. We thus wish to call for a critical reappraisal of what can be inferred from genetic or genomic data. The histories inferred by methods ignoring structure represent a first approximation but they are unlikely to provide us with the information we need to better understand the recent evolutionary history of humans or other species. It is difficult to imagine that humans have been one single panmictic population whose size has changed over the last few million years (*i.e.* since the appearance of the *Homo* genus) . This does not minimize the achievement of the Li and Durbin (2011) study, but it does question how inference from genetic data are sometimes presented and interpreted.

3.4.4 Perspectives

We focused throughout this study on T_2, the time to the most recent common ancestor for a sample of size two. For larger samples we can define T_k as the time during which there are k lineages. It would be important to determine whether, for structured models, the IICR estimated from the distribution of T_k varies significantly with k. If that were the case, that would suggest that it is possible to separate structure from population size change with the distributions of T_k for various k values. The reason for this is that population size change models should generate identical IICR for all T_k distributions, since they should all correspond to the same (real) history of population size change. To our knowledge the distribution of T_k for $k > 2$ has not yet been derived for the n-island or other structured models (but see interesting studies such as Wakeley and Aliacar (2001); Wakeley (2001); Nielsen and Wakeley (2001)).

One simple solution to this question is to simulate genetic data under a structured model of interest and then compare the simulated T_k distributions under that model and the T_k distributions of the corresponding model of population size change identified using the T_2 distribution. Preliminary simulations suggest that the T_k distributions produce different IICRs, at least for some models of population structure. For instance, we predict that the analysis of human genomic data with the PSMC and with the MSMC should produce different curves under a model of population structure but identical ones for a model of population size change. This prediction can be tested by comparing the PSMC and MSMC curves of Li and Durbin (2011) and Schiffels and Durbin (2014), respectively. Visual inspection of the corresponding figures suggests indeed that they are different, and therefore that our model of population structure is a valid alternative. However, we stress that an independent study is required. Indeed, the history reconstructed by these methods with real data is not very precise and the two curves are not easily comparable because they are expected to provide poor estimates at different moments. Any difference between the two analyses should thus be evaluated and validated with simulations.

Finally, one underlying assumption of our study is that the coalescent represents a reasonable model for the genealogy of the genes sampled. Given that the coalescent is an approximation of the true gene genealogy, and that there are species for which the coalescent may not be the most appropriate model (Wakeley and Sargsyan, 2009) we should insist that our results can, at this stage, only be considered for coalescent-like genealogies. The development of similar approaches for other genealogical models would definitely be a very interesting avenue of research.

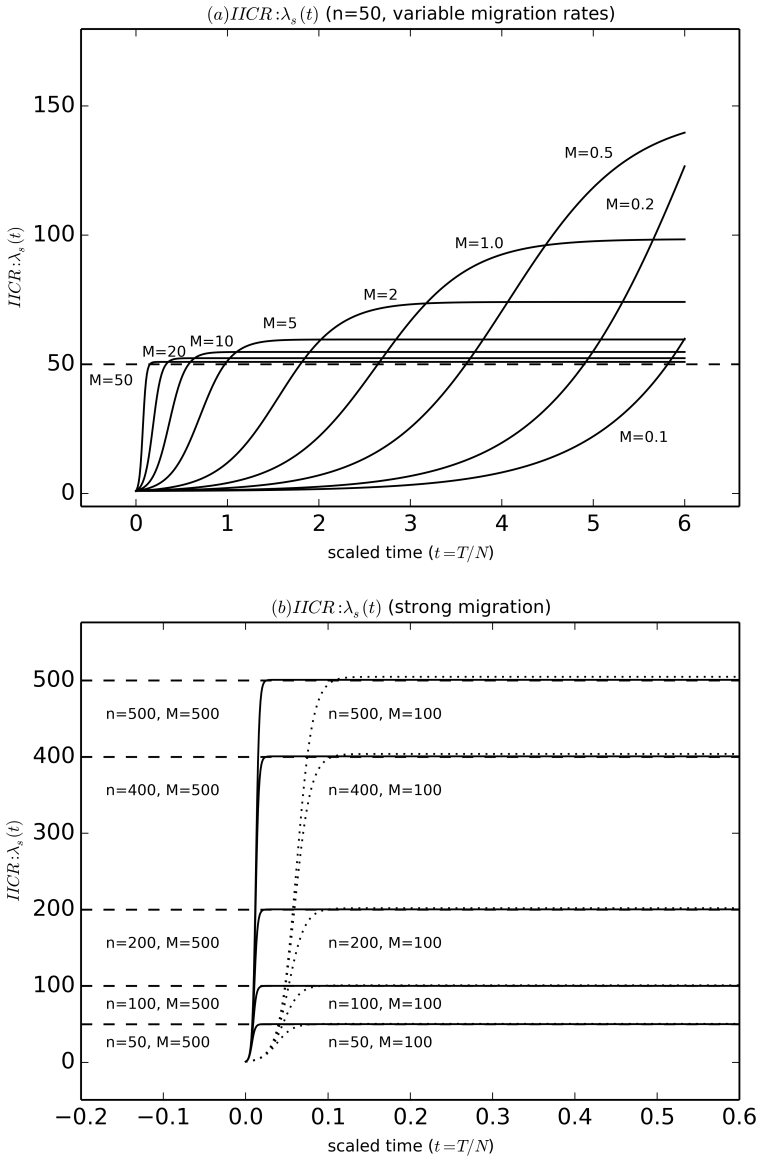


Figure 3.1: Inferred population size changes for n-island models with constant size. This figure shows $\lambda_s(t)$ for different values of M, the number of migrants, and n, the number of islands. In panel (a) we assumed an island model with $n = 50$, and varied M, the number of migrants between 0.1 and 50. In panel (b) we varied n between 50 and 500 and used two large values for M, namely 100 and 500. For both panels, the y axis is scaled by N and the horizontal dashed lines correspond to nN, the total population size. In all cases, $\lambda_s(t)$ identifies a population decrease.

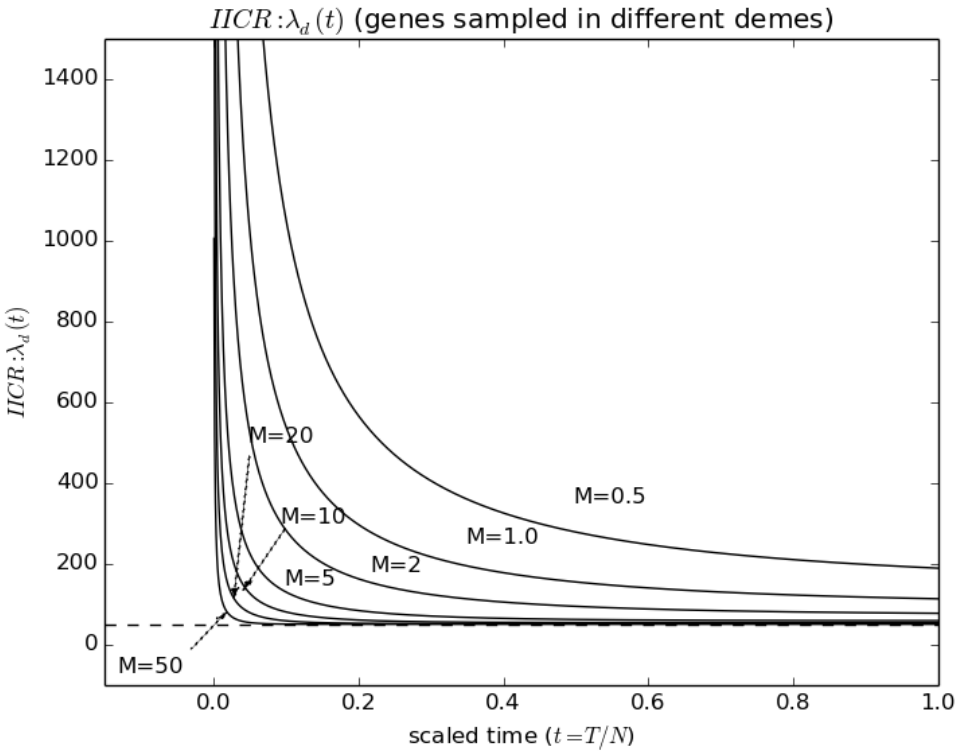


Figure 3.2: Inferred population size changes for n-island models and samples from different demes. This figure shows $\lambda_d(t)$ for different values of M, the number of migrants. The number of islands, was assumed to be $n = 50$. Samples come from different islands. In all cases, $\lambda_d(t)$ identifies a population increase.

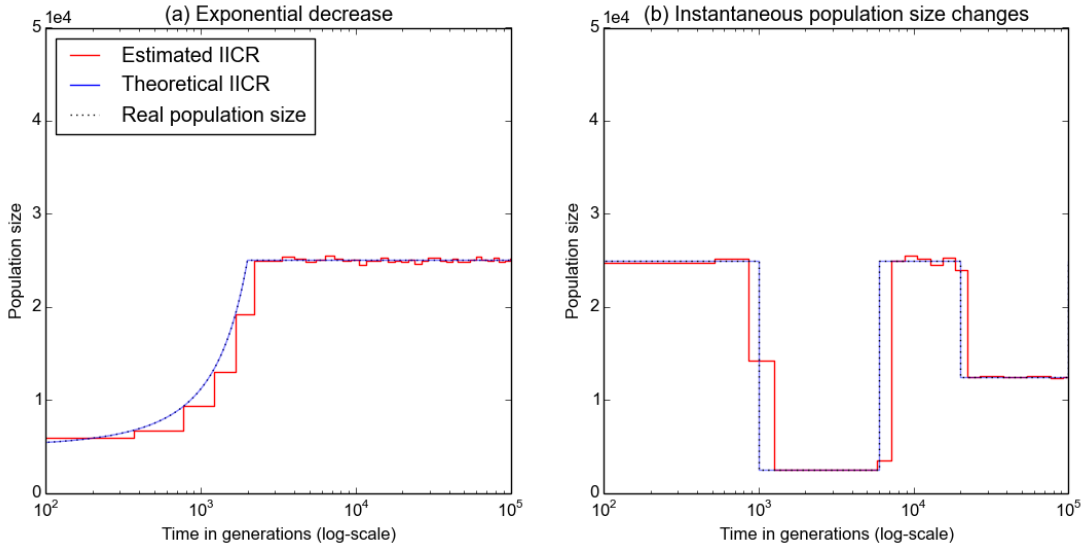


Figure 3.3: Inferred population size changes for populations without structure. For both panels the x-axis represents time in generations, whereas the y-axis represents population size in units of 10^4 diploids (an IICR of 0.5 corresponds to $500 * 10 = 5000$ diploid genomes). Panel (a) represents a panmictic population that experienced an exponential decrease from a previously constant size ancestral population. The solid blue line (theoretical IICR) was obtained using equation 3.4. The dashed line represents the simulated demographic history and corresponds to the total number of haploid genomes (the actual size). The stepwise red solid curve (estimated IICR) was obtained using the simulated T_2 values and equation 3.21. Panel (b) shows a history of stepwise population size changes. The color codes are identical to panel (a).

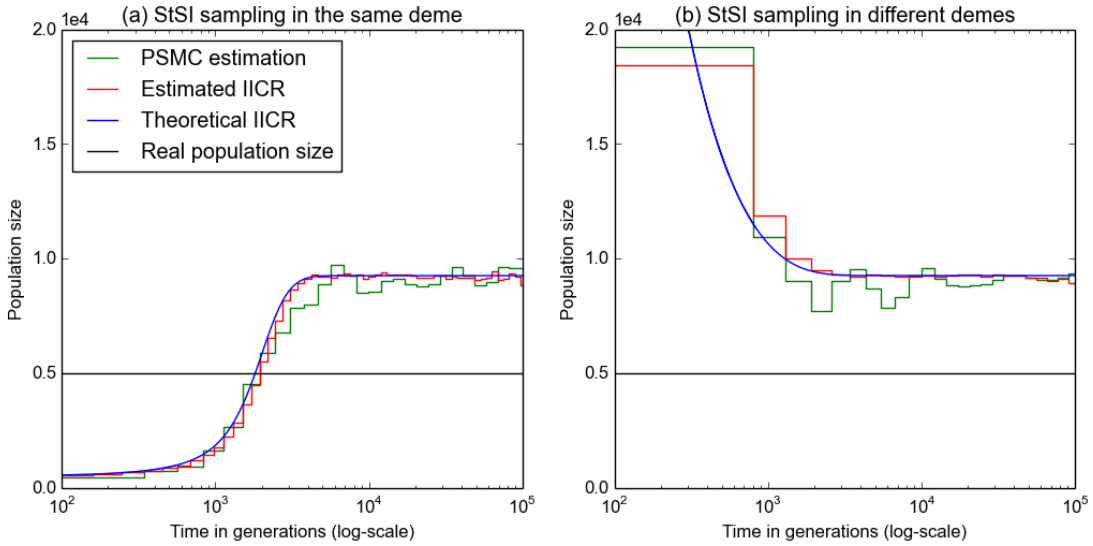


Figure 3.4: Inferred population size changes under population structure and two sampling schemes. This figure shows the predicted population size changes that will be inferred for an n-island model under the assumption that populations are not structured. For both panels the x-axis represents time in generations, whereas the y-axis represents real or inferred population size in units of 10^4 diploid genomes. We simulated an n-island model with $n = 10$ and $M = 1$ and computed the theoretical IICR using equation 3.4, and the estimated IICR using the simulated T_2 values and equation 3.21. The color codes are identical to Figure 3.3. The green solid lines represent the history inferred by the PSMC. Panel (a) shows the results when the two haploid genomes are sampled in the same deme. In panel (b) they come from different demes. The constant size of the metapopulation at $y = 0.5$ corresponds to $5,000$ diploid genomes or 10 islands of size 500 diploids.

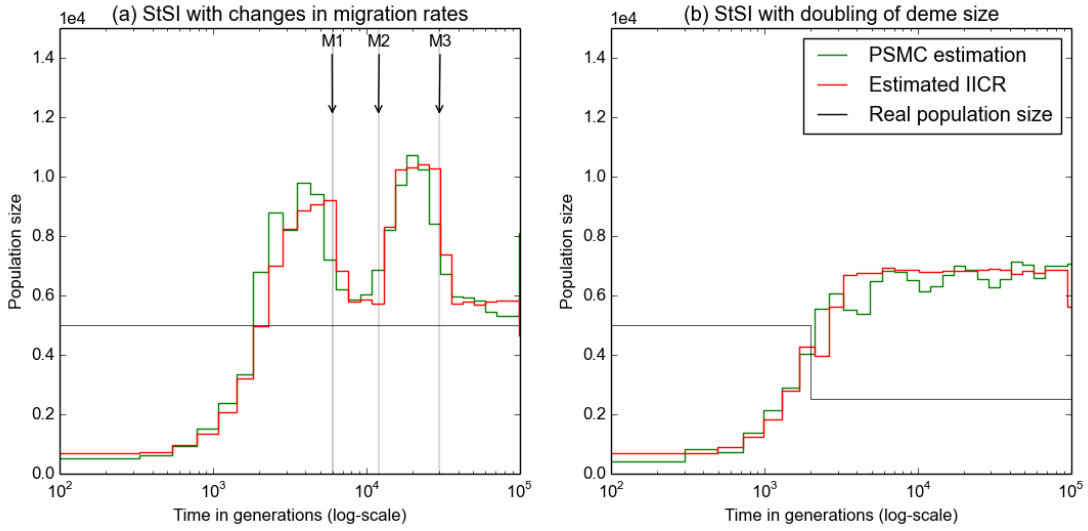


Figure 3.5: Inferred population size changes under population structure with changes in migration rates or deme size. The x-axis represents time in generations, whereas the y-axis represents real or inferred population size in units of 10^4 diploid genomes. Color codes are identical to figure 3.4. Data where simulated under an n-island model with $n = 10$. In panel (a) the population size was constant in size with each deme having a size $N = 1000$ haploid genomes (500 diploids) but three changes in migration rate occured at $T3 = 30,000$, $T2 = 12,000$, and $T1 = 6,000$ generations in the past. Before $T3$ the migration rate was $M3 = 5$. At $T3$ it changed to $M2 = 0.8$ and remained constant until $T2$, and then changed to $M1 = 5$ at $T1$. After that it remained at $M = 1$ until the present. In panel (b) all the demes doubled in size from 500 to 1,000 haploids (or 250 to 500 diploids) at $T = 2,000$ generations and migration was constant with $M = 1$.

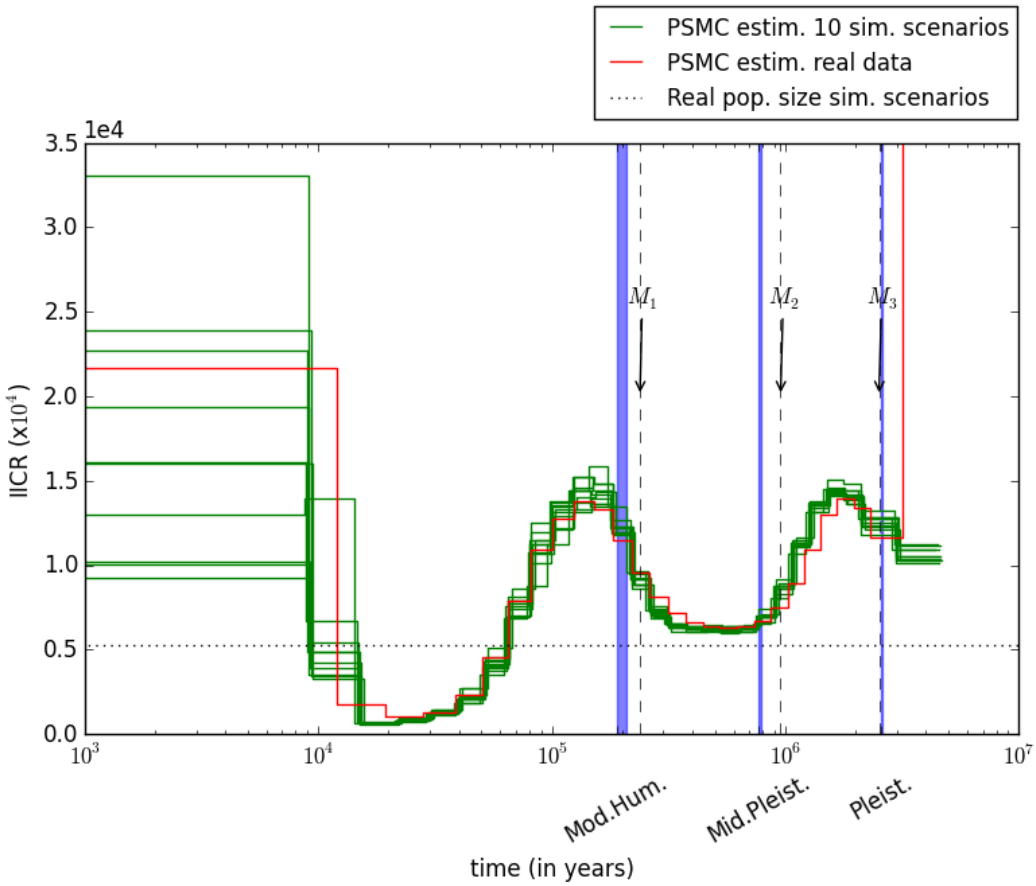


Figure 3.6: Human history with changes in migration rates. This figure shows, in red, the history of population size changes inferred by Li and Durbin from the complete diploid genome sequences of a Chinese male (YH) (Wang et al., 2008). The 10 green curves correspond to the IICR of ten independent replicates of the same demographic history involving three changes in migration rates. The x-axis represents time in years in a log scale, whereas the y-axis represents real or inferred population size in units of diploid genomes. The times at which these changes occur are represented by the vertical arrows at 2.52 MY ago, 0.95 MY ago and 0.24 MY ago. The blue shaded areas correspond to (i) the beginning of the Pleistocene (Pleist.) at 2.57-2.60 MY ago, (ii) the beginning of the Middle Pleistocene (Mid. Pleist.) at 0.77-0.79 MY ago, and (iii) the oldest known fossils of anatomically modern humans (AMH), at 195-198 KY ago. Following Li and Durbin (2011) we assumed that the mutation rate was $\mu = 2.5 \times 10^{-8}$ and that generation time was 25 years. We also kept their ratio between mutation and recombination rates. Each deme had a size of 530 diploids and the total number of haploid genomes was thus constant and equal to $10,600$.

Chapter 4

Detecting past demographic events in structured populations

In chapter 3 we gave the exact function of population size changes that will be inferred by a method assuming panmixia when applied to a population evolving under an n-island model (we named it IICR for *Inverse Instantaneous Coalescent Rate*). We have also shown that it is not possible, based on the distribution of T_2 (and consequently, based on any statistics computed from pairs of genes), to decide whether the population under study is structured or panmictic, because there will always be a panmictic population whose change in size can explain any distribution of T_2 values. As we illustrated in the examples, even if we were certain that the population is structured, the relation between the IICR (or the demographic history reconstructed by methods assuming panmixia) and past demographic events is not clear. When analysing a structured population with a method assuming panmixia, constant population size may lead to reconstructed histories with a clear signal of recent decrease or increase, depending on the sampling scheme. Sometimes, even a recent expansion in a population can be interpreted as a bottleneck. Even changes in gene flow can be interpreted as population size changes. The theory developed in the following chapter aims to clarify the relation between the IICR and past demographic events under a structured population.

The demographic history of a population has a strong influence on the genealogy of genetic samples. Consequently, if we assume some model for describing the evolution of a population over time, it is possible to depict the shape that coalescence trees will have. Then it is possible to compute the likelihoods of the parameters involved in the model, with respect to the trees or some observed statistics. Many population genetic studies are based somehow on this intuition and some of them manage to find the parameters that best explain observed data, by using a maximum likelihood (or a Bayesian) approach. However, unless the assumed model is

very simplistic, analytical expressions for the likelihood of the parameters based on observed data are very challenging to obtain. Besides, even minor modifications to the model in order to introduce a little bit of realism may highly increase the complexity of likelihood computations. In the following, we introduce the N-Island Markov Chain (NIMC), a framework based on the classical n-island model of Wright. The NIMC relies on a continuous Markov process to compute the distribution of coalescence times of two haploid individuals (or genes). By using this approach, it becomes possible to include past demographic events (like migration rate changes and population size changes) without any increase in the complexity of the likelihood expression. Moreover, this idea can be used as a way to detect population size changes beyond the confounding effect of population structure.

We start by constructing a Markov process that describes the evolution of two lineages backward in time under the n-island model, based on the ideas of the structured coalescent discussed in subsection 1.3.2. The Markovian property makes it easy to consider past demographic events like changes in migration rate and population size. Moreover, explicit expressions for the distribution function and the density of T_2 are derived, which makes it possible to trace the IICR in a precise way. We discuss some applications of the NIMC, especially that it is possible to accurately detect past changes in population size. We also propose a way to detect past demographic events based on the IICR inferred with methods assuming panmixia, and discuss how the NIMC framework can be directly applied to genomic data.

4.1 Coalescence times for a sample of size two in structured populations

The distribution of coalescence times in models that account for structure in the population has been a central point in many population genetics studies (Takahata, 1988; Notohara, 1990; Barton et al., 2002; Wakeley, 2001; Wilkins and Wakeley, 2002; Barton and Wilson, 1995). A very elegant extension of the coalescent was presented in Herbots (1994). This extension, named *Structured Coalescent*, is based on a continuous-time Markov chain and allows to compute explicitly the moment-generating function of the coalescence time of two genes under a wide range of models considering population structure (Herbots, 1994; Wilkinson-Herbots, 1998). The idea of a continuous-time Markov chain for tracking lineages backward in the time is also present in other works. For example, Wang and Hey (2010) used a Markov process to compute the coalescence time of two genes under the isolation-with-migration model (Nielsen and Wakeley, 2001). In this work a

three state continuous-time Markov chain was used to trace back two lineages until the MRCA (i.e. S_{11}: *both lineages are in subpopulation* 1, S_{22}: *both lineages are in subpopulation* 2 and S_{12}: *there is one lineage in subpopulation* 1 *and one in subpopulation* 2). Even though the authors did not provide analytical expressions for the distribution of the coalescence time, they gave a formula that can be approximated by using numerical integration methods. One year latter, Hobolth et al. (2011) proposed the idea of taking advantage of the continuous-time Markov chain representation to compute the distribution of coalescence times by mean of the matrix exponential, which led to very simple ways to write the expressions, that can be approximated numerically. Moreover, the authors noted that the matrix exponential framework can be extended to more than two populations and more than two genes. In the following we construct the N-Island Markov Chain (NIMC), a model allowing to detect past demographic events (changes in migration rate and change in population size) in a population evolving under the n-island model.

4.1.1 The N-Island Markov Chain: a continuous-time Markov process for the n-islands model

In the following paragraphs we present the N-Island Markov Chain (NIMC), a simplified version of the structured coalescent Herbots (1994) for the case of an n-island model. Just like the structured coalescent, the NIMC is a model for reconstructing the genealogy of genes back in time, from the present to the MRCA. Here, it is discussed a very simple case: we trace back just two lineages coming from two haploid individuals sampled in a population evolving under the n-island model. We will see that (given that we are considering a symmetrical gene flow between subpopulations), it is easy to include past demographic events into the model, specifically changes in the migration rates and the size of the total population.

We start by considering the classical n-island model of Wright (Wright, 1931). In this model, we have n islands (or *demes*) of constant size, connected by gene flow (figure 4.1.1 a.). The size of one island is considered equal to N haploid individuals. Given that all demes have the same size, the whole metapopulation size is nN haploid individuals. Islands are interconnected by the same amount of gene flow. Using the same notation as in subsection 1.3.2, backward migrations from deme i to deme j arrive at rate $M_{ij}/2$. This means that, going back in time, a lineage can migrate from deme i to deme j with rate $M_{ij}/2$. In the n-island model migration is symmetrical. So, backward migrations from any deme i to any other deme j arrive at the same rate. Let's define M such that, for any pair of demes (i, j), the rate of backward migrations from i to j is equal to $\frac{M}{2(n-1)}$. Consequently, if a lineage is in deme i, it migrates out of deme i at rate $\frac{M}{2}$.

Now, suppose that we take a sample of two haploid individuals (or two genes)

from the population and we trace back their lineages until they coalesce. Three different configurations (or states) are possible for the two lineages, when going from the present to the moment when they reach their common ancestor:

1. lineages are in the **s**ame subpopulation (state s or 1)
2. lineages are in **d**ifferent subpopulations (state d or 2)
3. lineages have **c**oalesced (state c or 3).

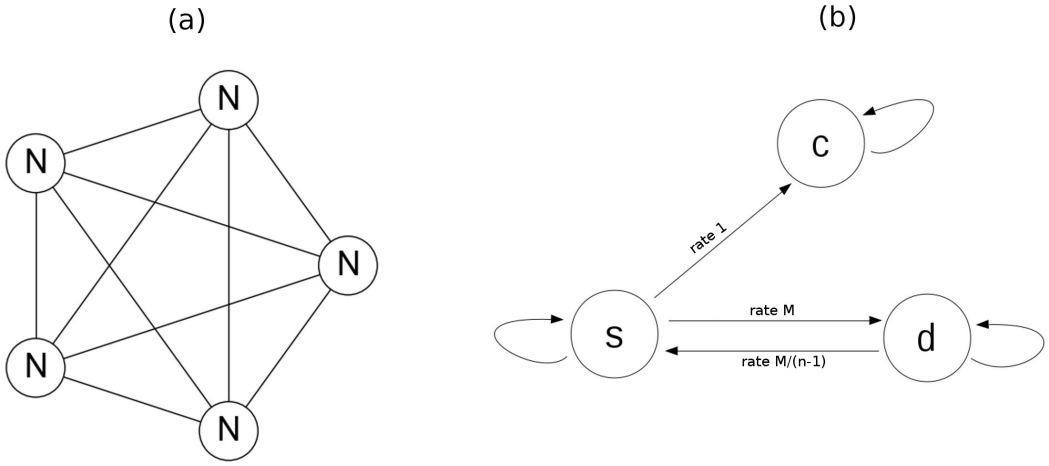


Figure 4.1: (a) n-island model for $n = 5$ islands. Each circle represents a deme of size N. All demes are connected to each other by symmetrical gene flow, represented by the edges. In this example the total number of haploid individuals is $5N$. (b) States of the continuous-time Markov process obtained when the lineages of two genes are traced back in time. Note that state c is an absorbing state.

In a similar way as in the structured coalescent (Herbots, 1994), we have a continuous-time Markov chain when tracing back lineages from the present until the MRCA. In this case, the process is simpler than that described by Herbots (1994) because, in the n-island model, migration rates are equal in all directions. Moreover, given that we are interested just in the coalescence of two haploid individuals, the process will stop when the two lineages coalesce. In other words, we can consider state c as an **absorbing state**. Let us now find the rates at which the process jumps from one state to another.

When lineages are in the same subpopulation (state s) a coalescence event may occur. Given that population size is constant, coalescence events inside the same subpopulation arise at rate one. Moreover, if the process is in state s, one of both

lineages may *migrate* to a different deme (note that this is a *backward migration* which occurs when the ancestor of one individual is in a different island). This occurs with rate $2 \times \frac{M}{2} = M$. On the other hand, when lineages are in different demes (state d), a coalescence event is not possible (lineages must be in the same subpopulation in order to coalesce). However lineages may migrate. Note that not all migrations make the process to jump from state d to state s. Assuming that one lineage is in subpopulation i and the other one is in subpopulation j (with $i \neq j$), a *backward migration* from deme i to deme j arises at rate $\frac{M}{2(n-1)}$ so as a *backward migration* from deme j to deme i. Thus, a migration event that implies a change from state d to state s arises at rate $2 \times \frac{M}{2(n-1)} = \frac{M}{n-1}$. Hence, the instantaneous rate matrix Q (the infinitesimal generator of the Markov process) is given by:

$$Q = \begin{pmatrix} -(M+1) & M & 1 \\ \frac{M}{n-1} & -\frac{M}{n-1} & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{4.1}$$

The values on the diagonal in the matrix Q are such that rows sum to zero. Once the infinitesimal generator is known, the *transition semigroup* of the Markov process can be computed by the matrix exponential:

$$P_t = e^{tQ}. \tag{4.2}$$

The cell (i, j) of the matrix P_t indicates the probability that the process is in state j at time t given that it was in state i at time zero. Thus, the probability that two genes sampled in the *same* subpopulation have reached their MRCA at time t can found in $P_t(1, 3)$ (first row, third column of matrix). In the same way, the probability that two genes sampled in *different* subpopulations have reached the MRCA at time t is equal to $P_t(2, 3)$. As in chapter 3, the coalescence time of two genes sampled in the same subpopulation will be denoted T_2^s, while that of two genes sampled in different populations will be called T_2^d. The cumulative distribution function (*cdf*) of these random variables can then be computed from the transition semigroup:

$$\begin{aligned} F_{T_2^s}(t) &= \mathbb{P}(T_2^s \leq t) = P_t(1, 3) \\ F_{T_2^d}(t) &= \mathbb{P}(T_2^d \leq t) = P_t(2, 3). \end{aligned} \tag{4.3}$$

Let S_0 and D_0 be the events considering that genes were sampled in the same subpopulation and different subpopulations respectively. Conditioning on S_0 and D_0, it is possible to write the distribution function of the time to reach the MRCA of two genes (T_2) by the formula of total probabilities:

$$\begin{aligned} F_{T_2}(t) &= \mathbb{P}(T_2 \leq t) = \mathbb{P}(T_2^s \leq t)\mathbb{P}(S_0) + \mathbb{P}(T_2^d \leq t)\mathbb{P}(D_0) \\ &= P_t(1,3)\mathbb{P}(S_0) + P_t(2,3)\mathbb{P}(D_0). \end{aligned} \tag{4.4}$$

Under the assumption that the sampling is done uniformly over the n islands, the distribution function becomes:

$$F_{T_2}(t) = \frac{1}{n}P_t(1,3) + \frac{n-1}{n}P_t(2,3). \tag{4.5}$$

The *density* of T_2 for these three cases can also be computed from the matrix P_t. This is done by using a classical property of the transition semigroup of a Markov process (Lemma 4.1.1), along with the particular form of the matrix Q.

Lemma 4.1.1. *Let P_t be the transition semigroup of a continuous-time Markov process with infinitesimal generator Q. Then:*

$$P_t' = P_t Q. \tag{4.6}$$

Proof. The transition semigroup can be computed from the infinitesimal generator by doing the matrix exponential:

$$P_t = e^{tQ} = \sum_{k=0}^{\infty} \frac{(tQ)^k}{k!}.$$

Taking the derivative, we obtain:

$$\Big(\sum_{k=0}^{\infty} \frac{(tQ)^k}{k!}\Big)' = \sum_{k=1}^{\infty} \frac{k(tQ)^{k-1}Q}{k!} = \Big(\sum_{k=0}^{\infty} \frac{(tQ)^k}{k!}\Big)Q = P_t Q.$$

□

By looking at the product of matrices P_tQ (and given that the third column of Q is the vector $(1,0,0)$) it can be noted that the third column of P_t' is equal to the first column of P_t. This implies that the *probability density function* (*pdf*) of the coalescence times of two genes can be directly recovered from the transition semigroup:

$$\begin{aligned} f_{T_2^s}(t) = F'_{T_2^s}(t) = P_t(1,1) \\ f_{T_2^d}(t) = F'_{T_2^d}(t) = P_t(2,1). \end{aligned} \tag{4.7}$$

By equation 4.5, the density of T_2 for an *n-island model* can be written as:

$$f_{T_2}(t) = F'_{T_2}(t) = \frac{1}{n}P_t(1,1) + \frac{n-1}{n}P_t(2,1). \tag{4.8}$$

4.1.2 An explicit expression for the transition semigroup

As sown above, evaluating the transition semigroup at time t only involves computing the matrix exponential. For relative small matrices, there is a variety of numerical methods available (Moler and Loan, 2003). In the very simple case considered here (i.e. a sample size of two from an n-island model), it is possible to obtain an analytical expression for the entries of the matrix P_t. In order to compute the matrix exponential given in equation 4.2 we should diagonalise the matrix Q. After some computations, it is possible to find that the eigenvalues of Q are:

$$\begin{aligned}\lambda_1 =& -\frac{1}{2}\Big(\gamma + M + 1 + \sqrt{\Delta}\Big),\\ \lambda_2 =& -\frac{1}{2}\Big(\gamma + M + 1 - \sqrt{\Delta}\Big),\\ \lambda_3 =& 0,\end{aligned} \tag{4.9}$$

with $\gamma = \frac{M}{n-1}$ and $\Delta = (\gamma + M + 1)^2 - 4\gamma$.
The corresponding eigenvectors are:

$$v_1 = \begin{pmatrix} 1 \\ \frac{\lambda_1 + M + 1}{M} \\ 0 \end{pmatrix}; \quad v_2 = \begin{pmatrix} 1 \\ \frac{\lambda_2 + M + 1}{M} \\ 0 \end{pmatrix}; \quad v_3 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}. \tag{4.10}$$

By denoting $\alpha = -\lambda_1$ and $\beta = -\lambda_2$ and inverting the matrix of the eigenvectors, it is possible to write the matrix Q as:

$$Q = ADA^{-1}, \tag{4.11}$$

where

$$A = \begin{pmatrix} 1 & 1 & 1 \\ \frac{M+1-\alpha}{M} & \frac{M+1-\beta}{M} & 1 \\ 0 & 0 & 1 \end{pmatrix}; \qquad D = \begin{pmatrix} -\alpha & 0 & 0 \\ 0 & -\beta & 0 \\ 0 & 0 & 0 \end{pmatrix};$$

$$A^{-1} = \begin{pmatrix} \frac{\beta - M - 1}{\beta - \alpha} & \frac{M}{\beta - \alpha} & \frac{1-\beta}{\beta - \alpha} \\ \frac{M+1-\alpha}{\beta - \alpha} & -\frac{M}{\beta - \alpha} & \frac{\alpha - 1}{\beta - \alpha} \\ 0 & 0 & 1 \end{pmatrix}.$$

Finally, observing that $\alpha\beta = \gamma$ and $\alpha + \beta = \gamma + M + 1$, and defining:

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

the transition semigroup can be computed explicitly:

$$P_t = \begin{pmatrix} ae^{-\alpha t} + (1-a)e^{-\beta t} & \frac{M}{\beta-\alpha}(e^{-\alpha t} - e^{-\beta t}) & 1 - \frac{a}{\alpha}e^{-\alpha t} - \frac{1-a}{\beta}e^{-\beta t} \\ ce^{-\alpha t} - ce^{-\beta t} & \frac{\beta-\gamma}{\beta-\alpha}e^{-\alpha t} - \frac{\alpha-\gamma}{\beta-\alpha}e^{-\beta t} & 1 - \frac{c}{\alpha}e^{-\alpha t} + \frac{c}{\beta}e^{-\beta t} \\ 0 & 0 & 1 \end{pmatrix} \tag{4.12}$$

Note that explicit expressions for $f_{T_2^s}$ and $f_{T_2^d}$ that can be found in $P_t(1,1)$ and $P_t(2,1)$ are in perfect agreement with previous theoretical works (Herbots, 1994; Mazet et al., 2015b,a). A general explanation for this fact will be given in below.

4.1.3 Incorporating past demographic events to the n-island model

The principal advantage of the NIMC framework is that we can use the semigroup property in a convenient way, in order to "restart" the process at any time. The semigroup property states that:

$$P_{t+u} = P_t P_u, \quad \forall t, u > 0. \tag{4.13}$$

From the perspective of the NIMC, this means that the probability of going from one state to another in time $t+u$ can be calculated by multiplying the infinitesimal generator evaluated at t (P_t) by the infinitesimal generator evaluated at u (P_u), and looking at the corresponding entry of the resulting matrix. Besides, the distribution function of T_2^s can be written by using the law of total probabilities: if S_t indicates that the system is in state s at time t, D_t that the system is in state d at time t and C_t that the system is in state c at time t, we have:

$$\mathbb{P}(T_2^s \leq t+u) = \mathbb{P}(S_t)\mathbb{P}(T_2^s \leq u) + \mathbb{P}(D_t)\mathbb{P}(T_2^d \leq u) + \mathbb{P}(C_t).$$

Replacing by the corresponding entries of the matrix P_t this can be written as:

$$\begin{aligned} P_{t+u}(1,3) =& \mathbb{P}(T_2^s \leq t+u) \\ =& P_t(1,1)P_u(1,3) + P_t(1,2)P_u(2,3) + P_t(1,3)P_u(3,3) \\ =& \Big(P_t P_u\Big)(1,3). \end{aligned} \tag{4.14}$$

If the matrix Q remains constant over the time, the process is a time-homogeneous Markov process. When we use two different matrices (say Q^0 in $[0,T[$ and Q^1 in $[T,+\infty[$), the process becomes time-dependent. However, equation 4.14 suggests that we could "restart" the process after a change in the Q-matrix, conditioning on where the lineages were in the instant of the change (i.e. at time T).

Changing migration rates Consider a population that evolves under an n-island model with a change in gene flow at some point in the past (let's say at $t = T$). This means that gene flow is equal to M_0 between the present and time $t = T$ (going from the present to the past) and it is equal to M_1 from time T to $+\infty$. Consequently, there will be two different Q matrices with the corresponding transition semigroups. Let P_t^0 be the transition semigroup corresponding to M_0 and P_t^1 that corresponding to M_1. Using the same reasoning as in equation 4.14, the *cdf* of the coalescence time of two genes sampled in the same island, under this new model can be computed as:

$$F_{T_2^s}(t) = \begin{cases} P_t^0(1,3), & \text{if } t \leq T \\ P_T^0(1,1)P_{t-T}^1(1,3) + P_T^0(1,2)P_{t-T}^1(2,3) + P_T^0(1,3) & \text{otherwise.} \end{cases} \tag{4.15}$$

In the same way, the distribution of coalescence times when genes are sampled in different islands is given by:

$$F_{T_2^d}(t) = \begin{cases} P_t^0(2,3), & \text{if } t \leq T \\ P_T^0(2,1)P_{t-T}^1(1,3) + P_T^0(2,2)P_{t-T}^1(2,3) + P_T^0(2,3) & \text{otherwise.} \end{cases} \tag{4.16}$$

More generally, if $\tilde{P}_t$ is the corresponding transition semigroup of the model described above, then:

$$\tilde{P}_t = \begin{cases} P_t, & \text{if } t \leq T \\ P_T^0 P_{t-T}^1 & \text{otherwise.} \end{cases} \tag{4.17}$$

The model can be extended in order to consider different values of gene flow at different times. Let

$$0 = t_0 < t_1 < ... < t_n < t_{n+1} = +\infty.$$

Assume for each interval $[t_i, t_{i+1})$ the gene flow is constant and equal to M_i. For each interval we have the corresponding infinitesimal generator (Q_i) and the transition semigroup (P_t^i). If $\tilde{P}_t$ is the infinitesimal generator of an n-island model with changes in the migration rate as described here, then we have:

$$\tilde{P}_t = \left(\prod_{i=0}^{k-1} P_{t_{i+1}-t_i}^i \right) P_{t-t_k}^k, \qquad (k = \max\{i | t_i < t\}). \tag{4.18}$$

The *cdf* of T_2^s and T_2^d under an n-island model with changes in gene flow can then be computed from $\tilde{P}_t$ in the same way as in equation 4.3. Regarding the probability density function (*pdf*) of T_2^s and T_2^d it can be noted that:

$$\tilde{P_t}' = \tilde{P}_t Q_k \qquad (k = \max\{i | t_i < t\}). \tag{4.19}$$

This and the fact that for any k the third column of Q_k is the vector $(1, 0, 0)$ allow to recover the *pdf* of T_2^s and T_2^d from $\tilde{P}_t$ in the same manner as in equation 4.7.

Changing population size Let's now consider that at each time t_i defined above, the size of the entire population changes by a factor of λ_i and remains constant inside $[t_i, t_{i+1})$. Consequently, the size of each island will be multiplied by the same factor λ_i. In this new scenario, with M_i and λ_i being the gene flow and population size change function on interval $[t_i, t_{i+1})$ respectively, the infinitesimal generator for each interval is:

$$Q_i = \begin{pmatrix} -(M_i + \lambda_i) & M_i & \lambda_i \\ \frac{M_i}{n-1} & -\frac{M_i}{n-1} & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{4.20}$$

and the corresponding transition semigroup can be computed as usual:

$$P_t^i = e^{tQ_i}. \tag{4.21}$$

For the matrix Q_i, we can have analogous results as in equations 4.9-4.11. The eigenvalues of the matrix Q_i are:

$$\begin{aligned} x_1^i &= -\frac{1}{2}\Big(\gamma_i + M_i + \lambda_i + \sqrt{\Delta}\Big) \\ x_2^i &= -\frac{1}{2}\Big(\gamma_i + M_i + \lambda_i - \sqrt{\Delta}\Big) \\ x_3^i &= 0, \end{aligned} \tag{4.22}$$

with $\gamma_i = \frac{M_i}{n-1}$ and $\Delta = (\gamma_i + M_i + \lambda_i)^2 - 4\lambda_i\gamma_i$.
The corresponding eigenvectors are:

$$v_1^i = \begin{pmatrix} 1 \\ \frac{x_1^i + M_i + \lambda_i}{M_i} \\ 0 \end{pmatrix}; \quad v_2^i = \begin{pmatrix} 1 \\ \frac{x_2^i + M_i + \lambda_i}{M_i} \\ 0 \end{pmatrix}; \quad v_3^i = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}. \tag{4.23}$$

Denoting $\alpha_i = -x_1^i$ and $\beta_i = -x_2^i$ and inverting the matrix of the eigenvectors, it is possible to write the matrix Q_i as:

$$Q_i = A_i D_i A_i^{-1}, \tag{4.24}$$

where

$$A_i = \begin{pmatrix} 1 & 1 & 1 \\ \frac{M_i+\lambda_i-\alpha_i}{M_i} & \frac{M_i+\lambda_i-\beta_i}{M_i} & 1 \\ 0 & 0 & 1 \end{pmatrix}; \qquad D_i = \begin{pmatrix} -\alpha_i & 0 & 0 \\ 0 & -\beta_i & 0 \\ 0 & 0 & 0 \end{pmatrix};$$

$$A_i^{-1} = \begin{pmatrix} \frac{\beta_i-M_i-\lambda_i}{\beta_i-\alpha_i} & \frac{M_i}{\beta_i-\alpha_i} & \frac{\lambda_i-\beta_i}{\beta_i-\alpha_i} \\ \frac{M_i+\lambda_i-\alpha_i}{\beta_i-\alpha_i} & -\frac{M_i}{\beta_i-\alpha_i} & \frac{\alpha_i-\lambda_i}{\beta_i-\alpha_i} \\ 0 & 0 & 1 \end{pmatrix}.$$

Observing that $\alpha_i\beta_i = \lambda_i\gamma_i$ and $\alpha_i + \beta_i = \gamma_i + M_i + \lambda_i$, and defining:

$$a_i = \tfrac{\gamma_i-\alpha_i}{\beta_i-\alpha_i}; \quad c_i = \tfrac{\alpha_i\beta_i}{\beta_i-\alpha_i},$$

the transition semigroup can be computed explicitly:

$$P_t^i = A_i e^{D_i} A_i^{-1} =$$
$$\begin{pmatrix} a_i e^{-\alpha_i t} + (1-a_i)e^{-\beta_i t} & \frac{M_i}{\beta_i-\alpha_i}(e^{-\alpha_i t} - e^{-\beta_i t}) & 1 - \lambda_i(\frac{a_i}{\alpha_i}e^{-\alpha_i t} + \frac{1-a_i}{\beta_i}e^{-\beta_i t}) \\ \frac{c_i}{\lambda_i}(e^{-\alpha_i t} - e^{-\beta_i t}) & \frac{\beta_i-\gamma_i}{\beta_i-\alpha_i}e^{-\alpha_i t} - \frac{\alpha_i-\gamma_i}{\beta_i-\alpha_i}e^{-\beta_i t} & 1 - \frac{c_i}{\alpha_i}e^{-\alpha_i t} + \frac{c_i}{\beta_i}e^{-\beta_i t} \\ 0 & 0 & 1 \end{pmatrix}. \tag{4.25}$$

As in equation 4.3, the *cdf* of T_2^s and T_2^s can be computed from P_t^i. Besides, by lemma 4.1.1 we have $(P_t^i)' = P_t^i Q_i$. Looking at the form of Q_i, it is easy to see that:

$$\begin{aligned} P_t^i(1,3)' &= \lambda_i P_t^i(1,1) \\ P_t^i(1,2)' &= \lambda_i P_t^i(1,2), \end{aligned} \tag{4.26}$$

which implies that the *pdf* of T_2^s and T_2^d can be computed as well from P_t^i.

The transition semigroup (called $\tilde{P}_t$) associated to an n-island model with changes in gene flows and population sizes as described above, can be computed as in equation 4.18. Also, equation 4.19 allows to evaluate its derivative at t. If we consider now T_2^s (T_2^d), the coalescence times of two genes sampled in the same island (different islands) under an n-island model with these modifications, then:

$$\begin{aligned} F_{T_2^s}(t) &= \tilde{P}_t(1,3) \\ F_{T_2^d}(t) &= \tilde{P}_t(2,3) \end{aligned} \tag{4.27}$$

and

$$\begin{aligned} f_{T_2^s}(t) &= F'_{T_2^s}(t) = \lambda_k \tilde{P}_t(1,1) \\ f_{T_2^d}(t) &= F'_{T_2^d}(t) = \lambda_k \tilde{P}_t(2,1) \qquad (k = \max\{i | t_i < t\}). \end{aligned} \tag{4.28}$$

In summary, the N-Island Markov Chain allows to compute the *cdf* and the *pdf* of the coalescence time of two genes sampled in a population evolving under an n-island model with changes in the gene flow and the metapopulation size over the time. The NIMC model is based on the transition semigroup of the Markov process that describes the history of two lineages back to the MRCA. The *cdf* and the *pdf* can be evaluated precisely by using explicit expressions, and in a very efficient way given that computations involve only the product of 3×3 matrices. Moreover, given that we have an exact expression for the *cdf* as well as the *pdf* of T_2^s and T_2^d, it is possible to compute the IICR by the formulas given below. This could be useful in order to predict the demographic history that will be reconstructed by methods like PSMC (Li and Durbin, 2011) when they are applied to structured populations. In practice, the main interest of the framework presented here is that it makes possible to detect past demographic events on structured populations, which is beyond the scope of actual methods. To our knowledge, most of the methods proposed for reconstructing population size changes through time are based on the panmictic hypothesis and therefore are sensible to gene flow and structure, which lead them to detect fake signals of bottleneck or expansions (Chikhi et al., 2010; Heller et al., 2013; Mazet et al., 2015a). The framework proposed here is able to overcome these limitations. In the following we expose how to use this framework to infer past demographic events from real data beyond the confounding effects of structure.

4.2 Applications of the NIMC framework

In order to have numerical values of the *cdf* and the *pdf* of T_2 for the scenarios under study, we have developed an implementation of the N-Island Markov Chain, using the python programming language. We carried out a validation of this implementation by comparing the values the *pdf* of T_2 with empirical distributions obtained from simulations using the software *ms* (Hudson, 2002). Details concerning these validations can be found in Annexes. The software is available at https://github.com/willyrv/nimc. In the following, we discuss some possible applications of the NIMC framework.

4.2.1 Detecting a bottleneck beyond the confounding effects of population structure

Perhaps the main advantage of the NIMC framework is that it gives a way to disentangle the effects of population structure when reconstructing the demographic history. It is well known that methods based on the assumption that the population is panmictic, are likely to find past population size changes when the population is structured, even if the population size has remained constant (Heller et al., 2013; Chikhi et al., 2010). In chapter 3 we gave the precise demographic history that methods assuming panmixia will reconstruct when applied to a population evolving under an n-island model, based on the distribution of the coalescence time of two genes. We also show that this demographic history depends on whether the two genes were sampled from the same island or from different islands. In the following, we will focus on the case where the two genes were sampled from the same island, because it is the one that corresponds to sampling one diploid individual from the population.

To illustrate how it is possible to avoid the confounding effects of the structure using the NIMC framework, we compared the results of parameter estimation under two structured scenarios involving population structure (Figure 4.2). The first scenario was an n-island model with constant size, ten islands and a migration rate of one ($n = 10$ and $M = 1$, Figure 4.2 left). The second was the same n-island model with $n = 10$ and $M = 1$ but having a recent increase in the population size by a factor of 10 at time $T = 0.5$. Using the ms software (Hudson, 2002), we simulated a set of 10000 independent values of T_2^s (the coalescence time of two genes, sampled in the same island), under these two scenarios. Then, we applied a maximum likelihood estimation based on two different models. The first model was the one used in chapter 2 for finding maximum likelihood estimates of the *time* (T) when bottleneck occurred and the *ratio* (α) of the bottleneck. This model assumes that population is panmictic. The second model also uses maximum likelihood estimation strategy, but assuming that population is structured. In this case we used the density given by the NIMC framework. More precisely, we used the density of T_2^s that corresponds to the NIMC model with constant migration rate and one single population size change. The number of island and the migration rate were fixed so that $n = 10$ and $M = 1$.

We can see from Figure 4.3 (left panel) that the method assuming panmixia (the first model described in the above paragraph) finds a strong bottleneck which agrees with Chikhi et al. (2010) and Heller et al. (2013). This is not surprising because, as it was explained in chapter 3, the demographic history reconstructed by methods based on this hypothesis follows the corresponding IICR, which is not always correlated with real population size changes. However, the method based on the NIMC framework (the second model described in the above para-

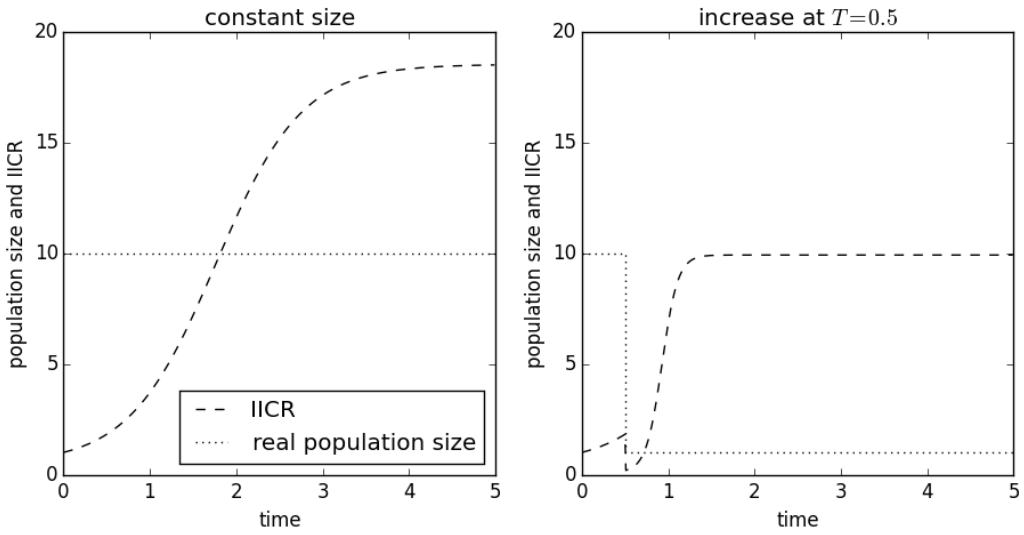


Figure 4.2: Two scenarios involving population structure and the corresponding IICR based on T_2^s (the coalescence time of two genes sampled in the same island). Left panel: an n-island model with $n = 10$ islands and migration rate $M = 1$. The population size and the migration rate are constant. The corresponding *ms*-command used for simulating the T_2^s values was: *ms 2 10000 -T -L -I 10 2 0 0 0 0 0 0 0 0 0 0 1.* Right panel: the same n-island model with $n = 10$ and $M = 1$ with an increase in population size of a factor 10 at time $T = 0.5$. The migration rate is constant. The corresponding *ms*-command used for simulating the T_2^s values was: *ms 2 10000 -T -L -I 10 2 0 0 0 0 0 0 0 0 0 0 1 -eN 0.25 0.1.* Note that, in the right panel, the population size as well as the IICR have been scaled with respect to the size of one island.

graph) finds that the ratio of the bottleneck is almost one (Figure 4.3, right panel), which corresponds better to the real population history in terms of population size changes.

As it has been described in chapter 3, the effects of structure in methods assuming a panmictic population can be so strong that a recent increase in population size may be unnoticed. Going back in time, we can see from Figure 4.2 (right panel) that the magnitude of the decrease in the IICR, caused by a recent increase by a factor of 10 in the population size, is small compared to the increase caused by the population structure. By consequence, when we try to fit a bottleneck to the data, based on a model assuming panmixia, we find a population size change in the opposite direction, that is, a decrease when actually the population has increased in size by a factor of 10 (Figure 4.4 left panel). Conversely, using a model based on the NIMC framework makes it possible to fit the bottleneck in the correct

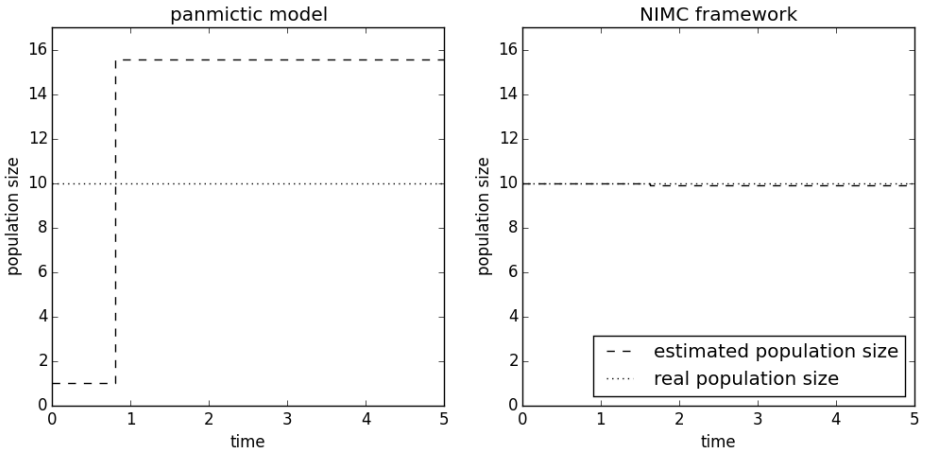


Figure 4.3: Spurious signals of population size changes found by methods assuming panmixia (left panel) disappear when inference is done under the NIMC framework (right panel).

direction, which corresponds with the real population size change (Figure 4.4 right panel). Applying a model assuming panmixia to a population which is structured, leads to find a decrease in population size by a factor close to 10, when actually, the population size has increased in size by a factor of 10. By contrast, when we analyse the same population using the NIMC framework, the inferred population size change is in agreement with the real demographic history.

Inferring the demographic history under the assumption of panmixia may lead to results that are strikingly different from those when structure is incorporated to the model. In many real scenarios, it is not very realistic to assume that population is panmictic. As a consequence, the demographic history reconstructed by a method based on this assumption, may not correctly reflect the changes in the population size. This confounding effect of population structure can be removed by using models that intrinsically incorporate structure to describe the evolution of populations. In this direction, the NIMC framework offers a very simple solution that can be incorporated to many inference methods in a relative simple way.

4.2.2 Using the transition semigroup for computing the instantaneous coalescence rate

Another interesting application of the NIMC framework is the possibility of tracing the IICR in a precise way. This implies that we will be able to predict the demographic history that any method based on the assumption that population is panmictic will find, when it is applied to a population evolving under any of

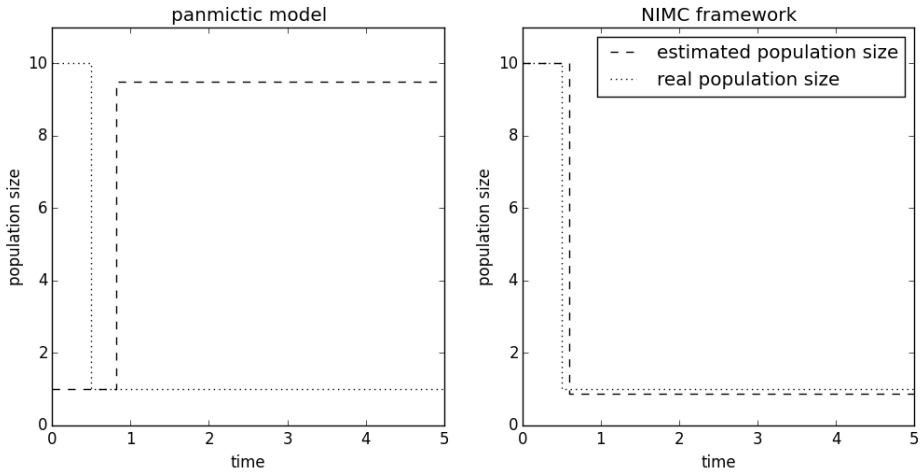


Figure 4.4: The problem of having bottleneck signals in the opposite direction of real bottlenecks when the population is structured, can be solved by incorporating population structure to the model.

the scenarios considered in the NIMC framework. Moreover, it is very efficient to compute IICR given that the evaluation of the *pdf* only involves the product of 3×3 matrices.

The relation between the distribution of coalescence times and the changes in population size through time is well known (see Tavaré (2004) for a detailed review) and many population genetic studies describe how it can be used to infer the demographic history (Mazet et al., 2015b; Pybus et al., 2000; Strimmer and Pybus, 2001). In order to detect past changes in population size, it is common to look for patterns in present days data that are somehow in relation with coalescence times. These patterns can be either differences between independent sequences of DNA (Drummond et al., 2005; Mazet et al., 2015b), microsatellites (Beaumont, 1999; Nikolic and Chevalet, 2014), allele frequency spectrum (Liu and Fu, 2015) or, more recently, full DNA sequences (Li and Durbin, 2011; Sheehan et al., 2013; Schiffels and Durbin, 2014; MacLeod et al., 2013). However, almost all the studies trying to estimate population size variations over the time consider an isolated panmictic population and neglect any kind of structure and gene flow. This may be a very limiting simplification when applying the model to real data, given that populations are structured to some extent in nature. Moreover, different models considering demographic changes and geographical structure can be explained equally well by the same class of genealogies (Nielsen and Beaumont, 2009). Besides, it has been shown that neglecting structure and gene flow may lead to detect fake bottleneck or expansion signals (Chikhi et al., 2010; Heller

et al., 2013) or even a bottleneck when the population size has in reality increased (Mazet et al., 2015a). In this sense, Mazet et al. (2015a) pointed out that the function of population size change inferred by most of the methods reconstructing past demography (denoted λ) is rather the *inverse* of the *coalescence rate* function that changes through time depending on the parameters of the assumed model. In Mazet et al. (2015a) authors noted that this λ function corresponds perfectly to population size changes in a panmictic population but in the general case, it is a coalescence rate function (called *IICR* or *Inverse Instantaneous Coalescence Rate*). Confounding the *IICR* with a population size change function inferred from data coming from a structured population may lead to detect spurious population size changes. Furthermore, by using the relation between the coalescence rate function and the distribution of the coalescence times of two genes under an n-island model, Mazet et al. (2015a) obtained the "fake demographic history" that should be reconstructed by methods like the PSMC (Li and Durbin, 2011), showing that this history is strongly dependent on the sampling strategy. The inverse coalescence rate function is time-dependent and (assuming that *cdf* and *pdf* of T_2 are known) can be computed as:

$$\lambda(t) = \frac{1 - F_{T_2}(t)}{f_{T_2}(t)}. \tag{4.29}$$

By using the NIMC framework, we can compute the λ function (or *IICR*) for a wide variety of scenarios, which allows to predict the "demographic changes" that most methods will find when they are applied to a population which is structured. For example, in a classical n-island model, the IICR that corresponds to two genes sampled inside the same subpopulation (λ_s), as well as in different populations (λ_d), can be obtained from the transition semigroup of the NIMC as follows:

$$\lambda_s(t) = \frac{1 - P_t(1,3)}{P_t(1,1)} \quad \text{and} \quad \lambda_d(t) = \frac{1 - P_t(2,3)}{P_t(2,1)}. \tag{4.30}$$

A more theoretical case is when we assume that the sampling is done uniformly over the n islands. We have then:

$$\lambda(t) = \frac{1 - \left(\frac{1}{n}P_t(1,3) + \frac{n-1}{n}P_t(2,3)\right)}{\frac{1}{n}P_t(1,1) + \frac{n-1}{n}P_t(2,1)}. \tag{4.31}$$

The IICR of an isolation with migration model The plots of λ_s presented in chapter 3 (Figure 3.1) suggest that when migration rate is high, populations should behave like panmictic populations, as it is intuitively expected. Previous works have considered structured models with high rate of migration, also known

as *the strong-migration limit* (Nordborg, 2001; Notohara, 1993; Nagylaki, 1980). In these works, it has been proved that coalescence trees in structured scenarios with high migration are equivalent to standard coalescence trees, after a change in the time scale. Consequently, the larger the migration rate, the closer the *cdf* of T_2^s in a structured model to the corresponding *cdf* of T_2 in a panmictic model. As expected, this also holds for the NIMC (Figure 4.5). For example, we simulated a sample of 3×10^6 independent values of T_2 under a model considering a constant-size population. This sample was compared with the corresponding theoretical distribution of T_2^s given by the NIMC framework with $M = 10000$ and constant population size. We did a Kolmogorov-Smirnov test and the obtained *p-value* was 0.7491. This means that three millions of markers are not enough to distinguish a panmictic model from an n-island model with $M = 10000$.

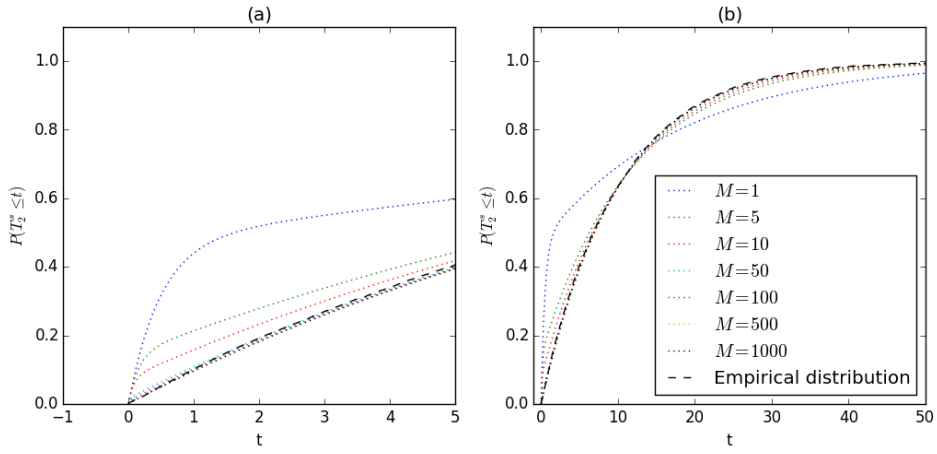


Figure 4.5: Comparisons between the empirical distribution of simulated values of T_2 under a panmictic model with constant population size, and the theoretical distribution of T_2^s of an n-island model with constant population size, for different values of the migration rate M. The size of the panmictic population is assumed to be equal to $10 \times N$, while the population evolving under the n-island model is composed by 10 subpopulations of size N. The values of the migration rate M are $1, 5, 10, 50, 100, 500$ and 1000. The left panel (a) and the right panel show the same data, with just a change in the time axis t.

It is possible to take advantage of this convergence to the standard coalescent, in order to have an approximation of the *cdf* and the *pdf* of T_2^s in models like the *isolation with migration* (Nielsen and Wakeley, 2001). This may be useful in order to compute the likelihood of parameters in these kind of models based on the observed data (for example, the time when population diverged, and the rate of migration after divergence), based on statistics depending on the distribution

of T_2^s like, for example, the number of pairwise differences. Moreover, having an expression for the *cdf* and the *pdf* allows to trace in a precise way the IICR that corresponds to a given model of population differentiation. As discussed in chapter 3, the IICR corresponds to the demographic history reconstructed by methods assuming that the population is panmictic. Consider that the population at time zero (the present) is subdivided into two subpopulations, connected with the same amount of gene flow, being $M = 1$. Going back in time, we can trace the IICR (i.e. λ_s) that corresponds to this scenario using the formula 4.30 (Figure 4.6 left). Note that, as t increases, the IICR approaches asymptotically to the line $y = 1/\beta$, where β is the value that was introduced in chapter 2 subsection 2.7.1 when doing the derivations for the density of T_2 under the n-island model. In correspondence with chapter 3 equation 3.17, we see that when t goes to infinity, the IICR tends to $1/\beta$. This limit value depends on the parameters of the model, in this case $n = 2$ and $M = 1$, and could be interpreted as the *structured ancestral effective size* (*saeffs*), going back in time, given that the IICR stabilises at this value when we go far enough in the past. Moreover, we have the following first order approximation:

Lemma 4.2.1. *Let n and M be the number of islands and the migration rate of an n-island model. Consider β as introduced in chapter 2, subsection 2.7.1. Then:*

$$\frac{1}{\beta} = n + \frac{1}{M}\frac{(n-1)^2}{n} + o\left(\frac{1}{M}\right), \tag{4.32}$$

where $\lim_{M\to\infty} Mo\left(\frac{1}{M}\right) = 0$.

Proof. We recall that:

$$\gamma = \frac{M}{n-1} \text{ and } \Delta = (1+\gamma)^2 - 4\gamma.$$

Following the definitions in subsection 2.7.1 we have:

$$\begin{aligned}\frac{1}{\beta} =& \frac{2}{1+n\gamma-\sqrt{\Delta}} = \frac{2(1+n\gamma+\sqrt{\Delta})}{(1+n\gamma)^2-\Delta} = \frac{1+n\gamma+\sqrt{\Delta}}{2\gamma} \\ =& \frac{n-1}{2M} + \frac{n}{2} + \frac{1}{2}\sqrt{\frac{\Delta}{\gamma^2}}.\end{aligned}$$

Replacing γ and taking n out of the square root, we get:

$$\begin{aligned}\frac{1}{2}\sqrt{\frac{\Delta}{\gamma^2}} =& \frac{n}{2}\left(1 + \frac{n-1}{M}(\frac{2}{n} - \frac{4}{n^2}) + \frac{(n-1)^2}{M^2n^2}\right)^{1/2} \\ =& \frac{n}{2} + \frac{n-1}{2M} - \frac{n-1}{nM} + o\left(\frac{1}{M}\right),\end{aligned}$$

and finally

$$\frac{1}{\beta} = n + \frac{1}{M}\frac{(n-1)^2}{n} + o\left(\frac{1}{M}\right).$$

□

Note that the *saeffs* is inversely proportional to the migration rate M. In other words, a low migration rate makes the ancestral population size look larger and a high value of M makes the ancestral population size look lower, from the perspective of the IICR. Consequently, methods assuming panmixia will detect a higher population size change when the migration rate is low. On the other hand, increasing the value of M not only decreases the *saeffs* but also makes the IICR converge quickly (see chapter 3 Figure 3.1). Moreover, for any value of M, the *saeffs* will always be higher than the number of islands n (this can be observed from the equations in Theorem 4.2.1) and it approaches asymptotically to n when M increases.

Using the NIMC framework, it is possible to approximate a model like the one proposed in Nielsen and Wakeley (2001) (with the difference that here the gene flow is symmetrical). In this model, the authors consider two populations which are descendent from a panmictic ancestral population. A model like the one in Nielsen and Wakeley (2001), can be approximated by an n-island model with two islands and one change in the migration rate M at some time T in the past. To that end, we consider a population which is subdivided into two colonies at the present, with migration rate $M = 1$. Going back in time, at time T, the migration rate changes from $M = 1$ to $M = 10000$. In Figure 4.6 we show some plots of the IICRs corresponding to the described scenario, for different values of T (the time when populations diverged). The considered values for the time when populations diverged were $T = 5, T = 2$ (left panel) and $T = 1, T = 0.5, T = 0.1$ (right panel). Note that, for a classical n-island model with $n = 2$ and $M = 1$, the IICR is almost constant and equal to the *saeffs* for $t > 3$ (Figure 4.6 a). Starting from the present ($t = 0$), the IICR increases from 1 to some value (that we have called the *saeffs*). Once the IICR stabilises close to this value, increasing the migration rate M makes the IICR decrease to a different *saeffs* which depends on the parameter n and the new value of M. Going back in time, the IICR remains almost constant and equal to this new *saeffs* after the increase of the value of M. As expected, the *saeffs* corresponding to high values of M in this scenario is close to 2. For low values of M, the IICR starts from one and increases until some *saeffs* which is higher than 2. If the change from M_0 to M_1 (with $M_0 < M_1$) arrives after the IICR exceeds the *saeffs* corresponding to M_0, we will observe a reduction of the IICR, going back in time (Figure 4.6 a). Consequently, if we interpret the demographic history reconstructed by a method assuming a panmictic population (now going

forward in time), we will detect a sudden increase in the population size at the time when the splitting occurred, followed by a period (long or short, depending on when the split occurred) when population was constant and then a decrease in a recent past. Thus, if the split is more ancient than the time when the IICR reaches the *saeffs*, we will observe a bump. On the other hand (Figure 4.6 b), if the split is more recent than the time when the IICR reaches the *saeffs*, we will observe a sudden bottleneck signal (going forward in time).

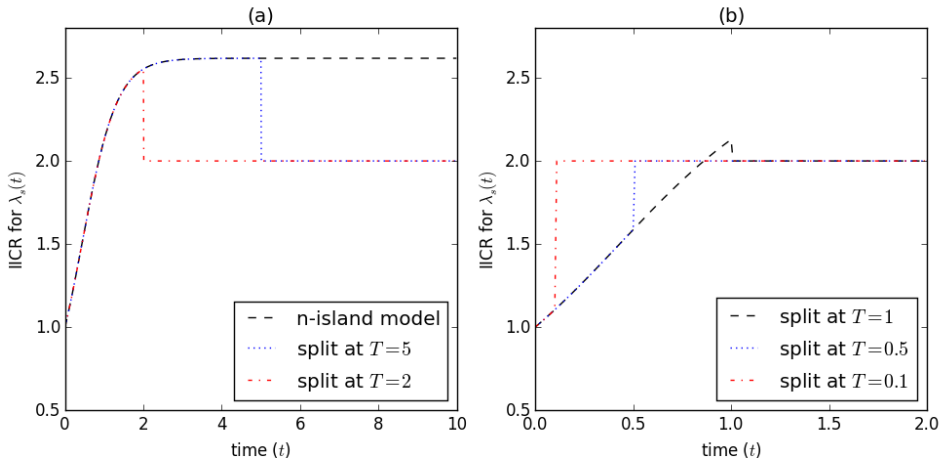


Figure 4.6: The IICR for an isolation-with-migration model for distinct values of T, the time when populations diverge. Panel (a): classical n-island model (or a model having an split older than $T = 10$), split at $T = 5$ and split at $T = 2$. Panel (b): the split occurs at $T = 1$, $T = 0.5$ and $T = 0.1$.

Signals of population size changes explained by an n-island model with changes in gene flow The above discussion gives some intuitive ideas on how changes in gene flow in an n-island model produces "fake" signals of population size changes. Going back in time, provided that the IICR has stabilised around the *saeffs*, an increase of gene flow will cause a decrease in the IICR. Conversely, if we reduce gene flow, the IICR is expected to increase, leading to an increase in the estimated population size, when we use a method assuming panmixia. This suggests that past population size changes found in many studies carried out by using methods based on panmixia (Li and Durbin, 2011; Prufer et al., 2014; Cahill et al., 2013; Zhou et al., 2013), could also be explained by scenarios of structured population with changes in gene flow and with absolutely no change in population size. To illustrate this, we applied the PSMC method (Li and Durbin, 2011) to one of the human sequences analysed in Li and Durbin (2011), we removed the most recent part of the inferred history and scaled it conveniently so the first value

(starting from the most recent value) is equal to one. Then, we proposed an n-island model with constant population size and three changes in migration rates at three different times, for which the IICR is surprisingly similar to the demographic history inferred by PSMC (Figure 4.7). This can be interpreted as follows. Consider a scenario of structured population for describing the evolution of the human species. Assume that humans have evolved under an n-island model with constant population size and that genetic flow between subpopulations has changed at three different times in the past. Based on the demographic history inferred by PSMC (or by any other method assuming panmixia), the described structured scenario could be indistinguishable from one scenario considering a panmictic population with changes in population size. We followed the same procedure for the Neanderthal genome analysed in Prufer et al. (2014) and we also found an n-island model with constant size and changes in gene flow, whose corresponding IICR is very close to the demographic history inferred by PSMC (Figure 4.7). In each case, the scenario having an IICR close to the inferred demographic history, was found by tracing the IICR corresponding to different scenarios and comparing them to the target demographic history. By using the intuition given above (increasing the migration rate makes the estimated population size to be lower, and decreasing the migration rate makes it to be higher), it was possible to find scenarios having an IICR very close to the demographic history inferred by PSMC which is based on a model assuming panmixia.

It is important to clarify that the objective of the above paragraph is not at all to propose a new way for interpreting the history of human evolution. We are convinced that a symmetrical n-island model (even including changes in population size and migration rate) is still a very simplistic model to describe the evolution of any species. It is also simplistic to consider that reproduction between individuals in a population occurs at random. For these reasons, the results obtained with any model should be interpreted with care, when analysing a real population. In any case, what is important here is that the NIMC makes available a wide range of alternative scenarios that worth considering when inferring the demographic history of populations.

On the other hand, there may be scenarios considering a panmictic population with changes in population size, for which it is not possible to find any n-island model with constant size and changes in gene flow, having an IICR close to the one of the panmictic population. In other words just changes in gene flow are not enough to interpret any signal of population size change. For example, a strong signal of bottleneck in the past cannot be explained by an n-island model with changes in gene flow, if the bottleneck is more ancient than the time when the IICR of an n-island model "stabilises" close to the *saeffs*. Beyond this time (going back in time), the IICR of an n-island model is never under n (the number of

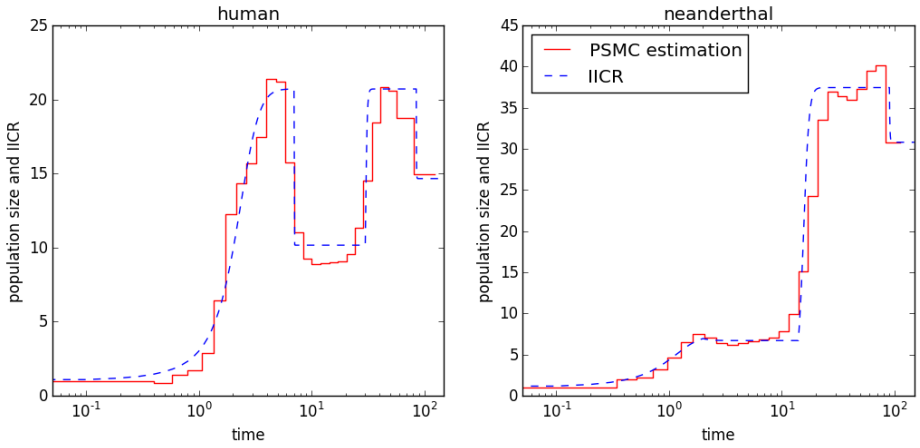


Figure 4.7: Searching for structured models having the same IICR that the proposed panmictic models with population size changes. In both cases, we consider an n-island model with constant population size and three changes in migration rate at three different times. Left panel: the demographic history reconstructed by PSMC in Li and Durbin (2011) from the DNA of one individual (labelled *CHN.A*) also corresponds to the IICR of an n-island model with 10 islands and constant population size. The changes in the migration rate are given (backward in time) by the vectors $T = (0, 7, 30, 85)$ and $M = (0.8, 50, 0.8, 1.8)$. Right panel: the history inferred by PSMC from a Neanderthal genome in Prufer et al. (2014) is close to the IICR of an n-island model with five islands ($n = 5$), constant population size and changes (back in time) given by $T = (0, 2, 14, 90)$ and $M = (1.5, 2, 0.12, 0.15)$.

islands), no matter how much the migration rate increased. This means that, beyond this time, a bottleneck making the population size to decrease under the value of n cannot be explained by an n-island model considering only changes in the migration rate.

4.3 Perspectives

4.3.1 Inferring parameters based on the IICR obtained from other methods

In Figure 4.7 we show that it is possible, given a curve representing the population size changes of a panmictic population, to find an n-island model with changes in migration rates, whose IICR closely corresponds to the given population size changes under panmixia. This suggests an indirect way to do inference of model parameters under the NIMC framework based on the demographic history (which is

not other than the IICR) inferred by any method assuming panmixia (like PSMC or MSMC). The inferred demographic history can be used as "target" curve to be approximated by the IICR of a model described by the NIMC framework. Changing the parameters of the NIMC model will make the corresponding IICR change. This can be done in a convenient way, until we get a model whose IICR is close to the given demographic history.

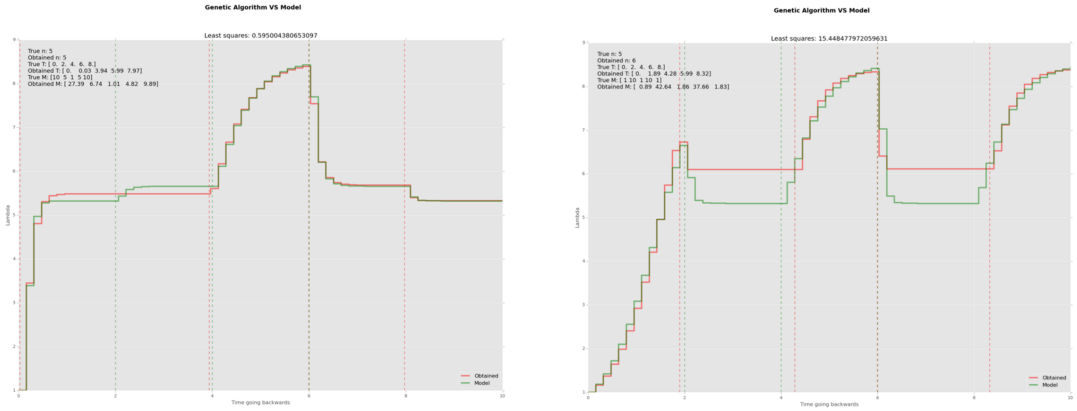


Figure 4.8: Demographic inference by a curve fitting process using the IICR of the NIMC.

We developed a python implementation of the described procedure. The software is available at https://github.com/MaxHalford/StSICMR-Inference. Basically the software tries to solve a curve fitting problem by using the IICR corresponding to different NIMC-based models, with different set of parameters (Figure 4.8). The objective is to find an n-island model with changes in gene flow and population size, whose corresponding IICR is as close as possible to a given curve (which is the "target"). The parameters to estimate are: the number of islands (n), the times when changes in gene flow or population size occur (a vector T), the values of migration rate in any time interval (a vector M) and the values of population size, starting by 1 at the present (a vector α). The target curve can be the one obtained by a PSMC analysis on real data. The method uses a genetic algorithm to minimise the distance between the IICR and the "target" curve, evaluated at some different points. Even if the parameter space is huge, the method manage to find models whose IICR is close to the "target" demographic history. However, from our experience, it is possible to get an IICR even closer to the one found by the minimisation algorithm, simply by visual comparisons using the notions discussed above. For example, both IICRs in Figure 4.7 where found very quickly by visual comparisons and intuitively changing the corresponding parameters.

4.3.2 Links with observable quantities

Having an analytical expression for the density of T_2 can be useful in order to estimate model parameters. Based on this density, Maximum Likelihood Estimation (MLE) strategies can be proposed to find (at least in theory) good estimations for the time when gene flow changed, as well as for the magnitude of this change. Likewise, a MLE strategy could be used to reconstruct population size changes in an n-island model, based on values of T_2. However, coalescence times are not observable quantities and even if some authors have published works allowing to estimate trees from genomic data (Pybus et al., 2000; Strimmer and Pybus, 2001; Drummond et al., 2005) it is still very challenging to reconstruct in a precise manner the genealogy based on the data. Other methods based on Hidden Markov Models and taking into account the recombination (Li and Durbin, 2011; Schiffels and Durbin, 2014; Sheehan et al., 2013) are shown to accurately infer the coalescence times of two locus along the chromosome, provided a previous set of time windows is defined. If a stepwise approximation to the *pdf* of T_2 is considered, this inferred values could be interpreted as observations which would make possible to estimate model parameters by MLE. Even though this idea worth a deep exploration, we prefer to link model parameters and genomic data in a more direct way. In the following, two approaches will be discussed: one based on the number of segregating sites of non recombining loci (Watterson, 1975; Mazet et al., 2015b) and the other based on a Hidden Markov Chain along the chromosomes, in the same way as Li and Durbin (2011).

Number of segregating sites It is possible to compute, conditional on the values of T_2, the distribution of several measures of molecular polymorphism. For example, under an infinite site mutation model with mutation rate θ, it is possible to compute the number of differences between pairs of non recombining sequences (N_d), as it was proposed in subsection 2.6.1. If we suppose that the coalescence time of two non recombining DNA sequence is equal to t, the distribution of N_d is a Poisson distribution with parameter 2θ. The number of difference between pairs of non recombining sequence can then be computed as:

$$\mathbb{P}(N_d = k|T_2 = t) = e^{-2t\theta}\frac{(2t\theta)^k}{k!}. \tag{4.33}$$

As we know the density of T_2, we can take the integral over all possible values of t:

$$\begin{aligned}\mathbb{P}(N_d = k) =& \int_0^{+\infty} \mathbb{P}(N_d = k | T_2 = t) f_{T_2}(t)\, dt \\ =& \frac{1}{k!} \int_0^{+\infty} e^{-2t\theta} (2t\theta)^k f_{T_2}(t)\, dt.\end{aligned} \tag{4.34}$$

This integral can be computed using numerical methods in a relatively efficient way, given that $f_{T_2}(t)$ can be evaluated at any t, using the explicit expression from the NIMC model or computing the matrix exponential numerically.

Constructing a Hidden Markov Chain over the genome In Li and Durbin (2011), authors proposed a new method (the PSMC) for estimating the population size at different time intervals in the past from a single diploid genome. This method is based on the Sequentially Markovian Coalescence (McVean and Cardin, 2005b) which is an approximation to the coalescent with recombination (Hudson, 1983). The PSMC uses a Hidden Markov Chain were the observations are given by a diploid sequence (actually there are only two possible observed states at each position inside the genome: homozygous or heterozygous). The hidden states are the coalescence times at each position. When moving along the genome, recombination events break the chromosomes, making a mosaic of distinct chunks of DNA, inherited from different ancestors. Consequently, the coalescence time of two genes at the left of a recombination event is different from the coalescence time at the right side of the recombination, given that they were not inherited from the same ancestor. This implies a change from one state to another. The transition probability between hidden states was explicitly computed by Li and Durbin (2011) in function of the population size at each time in the past (i.e. $\lambda(t)$). Thus, the PSMC model is able to compute the likelihood of a given demographic history, with respect to the observed data (the observed data is actually the full genome of a diploid individual). The method uses the Expectation Maximisation (EM) algorithm to find a demographic history who maximises the likelihood of the observed genome.

The PSMC method is based on the assumption of panmixia. As discussed in chapter 3, the demographic history inferred by this kind of methods is what we defined as IICR. This means that, if we consider an n-island model, instead of a panmictic population, the transition probabilities described in the PSMC model depend on the IICR. Moreover, in this chapter we propose a way to evaluate the IICR over time. Combining these two models, it is possible to compute the likelihood of the parameters of any model described by the NIMC, based on full diploid genomes. Then, the EM algorithm, or any other maximum likelihood strategy, could be used to find the values of the parameter that best explain the observed data.

4.3.3 Extending the NIMC to more than two genes

The same reasoning used to derive the transition semigroup of the Markov process, that arises when two lineages are traced back in time, under an n-island model, can be extended to three lineages. If we take three genes from the population and trace back their lineages, until the time of the first coalescent event (that we denote T_3), there are four possible states for the ancestral lineages at any time t between the present ($t = 0$) and T_3:

1. the three ancestral lineages are in the same deme;
2. exactly two among three remaining lineages are in the same deme;
3. all the three lineages are in different demes;
4. two of the three lineages have coalesced.

Given that we are interested in tracing the lineages on-ly until the time when the first coalescent event occurs, we consider that state 4 is *absorbing*. Assuming that the number of subpopulations is equal to n and that the rate at which one lineage migrates is equal to $M/2$, we proceed as in subsection 4.1.1 and get that the corresponding Q-matrix in this case is:

$$Q = \begin{pmatrix} -\frac{3M}{2} - 3 & \frac{3M}{2} & 0 & 3 \\ \frac{M}{2(n-1)} & -\frac{M(2n-3)}{2(n-1)} - 1 & \frac{M(n-2)}{n-1} & 1 \\ 0 & \frac{3M}{n-1} & -\frac{3M}{n-1} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{4.35}$$

Once the Q-matrix is known, we have, for example, the *cdf* of the time when the first coalescent event occurs, conditioned by sampling the three genes from the same deme. Denoting S, the event indicating that the three genes were sampled from the same deme and T_3^s, the time of the first coalescence event, we have:

$$\mathbb{P}(T_3^s \leq t) = \mathbb{P}(T_3 \leq t|S) = P_t(1,4) = e^{tQ}(1,4).$$

It is also possible to trace back three lineages until the MRCA. In this case we have 6 possible states for the Markov process:

1. the three ancestral lineages are in the same deme;
2. two lineages are in the same deme and one in a different deme;
3. all the three lineages are in different demes;

4. only two lineages have coalesced (which means that there are only two ancestral lineages left), and the two remaining ancestral lineages are in the same deme;

5. only two lineages have coalesced (which means that there are only two ancestral lineages left), and the two remaining ancestral lineages are in different demes;

6. the three lineages have coalesced.

The corresponding Q-matrix is given by:

$$Q_3 = \begin{pmatrix} -3(\frac{M}{2}+1) & \frac{3M}{2} & 0 & 3 & 0 & 0 \\ \frac{M}{2(n-1)} & -\frac{M(2n-3)}{2(n-1)} - 1 & \frac{M(n-2)}{n-1} & 0 & 1 & 0 \\ 0 & \frac{3M}{n-1} & -\frac{3M}{n-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & -(M+1) & M & 1 \\ 0 & 0 & 0 & \frac{M}{n-1} & -\frac{M}{n-1} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{4.36}$$

Remark: In the bottom right-hand corner of the matrix Q_3 we recognise the rate matrix Q, which describes the genealogical process for two lineages. The sixth state is an absorbing state and corresponds to the state when the three lineages found their MRCA.

Once the entries of the Q-matrix are computed, we can compute the transition semigroup by the matrix exponential. In the general case, analytical expressions for the transition semigroups can be hard to obtain but we can evaluate it at any value of t by computing the matrix exponential numerically (Moler and Loan, 2003). However, in practice it could be hard to compute the transition semigroup of the Markov process for larger numbers of genes given that, when the number of lineages increases, the number of states will grow rapidly, making the Q-matrix so big that computing the matrix exponential will become computationally intractable.

Extending the NIMC framework to samples larger than two, makes it possible to compute the probability of a given genealogy under an n-island model incorporating changes in gene flow and population size. This could be useful when trying to estimate parameters related to the demographic history. Moreover, the matrix Q_3 given above can be used to compute the joint density of (T_3, T_2). This joint distribution under an n-island model could be substantially different from the same joint distribution under a panmictic model with any function of population size changes (λ). If this is the case, then it would be possible to construct a method for distinguish an n-island model from a general scenario assuming panmictic population with any function of population size changes, using tree haploid individuals.

Conclusion

Les travaux présentés dans cette thèse sont le résultat d'un cheminement important. Mathématicien de formation j'ai commencé à travailler sur les données génomiques pour la première fois avec Simon Boitard à l'INRA ainsi qu'avec Olivier Mazet. Le but était de me familiariser avec une méthode qui venait d'être développée, le PSMC (Li and Durbin, 2011) et de l'appliquer à des données de séquençage de races de mouton. Lorsque l'idée de faire une thèse a commencé à prendre forme une des idées initiales était de développer une approche similaire mais pour plusieurs génomes. J'ai donc passé les premiers mois de ma thèse à essayer de comprendre tous les éléments nécessaires à l'implémentation du PSMC. J'ai aussi beaucoup travaillé avec Simona Gruséa sur les certains problèmes mathématiques. Par ailleurs lorsque j'avais commencé à discuter des résultats sur les moutons dans le cadre de réunions de travail avec O. Mazet, S. Oitard, S. Gruséa et L. Chikhi, ce dernier a mentionné le problème de la structure des populations. Il notait que que les races de moutons sont particulièrement structurées et que certains changements de tailles pourraient n'avoir jamais eu lieu mais correspondre à des changements de flux géniques. Cette idée qui fut très vite partagée par tous les membres impliqués dans ce travail a fait son chemin et m'a amené à changer de manière significative le centre de mon travail. Le fait que les méthodes de DiCal (Sheehan et al., 2013) et MSMC (Schiffels and Durbin, 2014) aient été publiées ont aussi facilité ce changement. Nous nous doutions bien entendu que d'autres groupes travaillaient sur ce problème (l'utilisation de génomes multiples pour reconstruire l'histoire des populations). Et notre intérêt pour la structure ne faisait que croître.

D'autant plus que la plupart des méthodes d'inférence démographique en génétique des populations sont centrées sur les changements de taille efficace. Ces changements sont devenus quasiment synonymes d'histoire démographique. comme nous l'avons dit à de nombreuses reprises, ces méthodes supposent que la population (en réalité, l'espèce) est panmictique, en négligeant l'existence d'une structure qui changerait le modèle de reproduction des individus. Dans cette thèse, nous nous sommes donc intéressé à des effets de la structure sur les changements de taille inférés par les méthodes basées sur l'hypothèse de panmixie.

Différents travaux publiés avant même que cette thèse n'ai commencé ont mon-

tré que, lorsqu'on analyse une population structurée avec une méthode qui suppose que la population est panmictique, certains changements inférés ne correspondent pas à des vrais changements de taille. Donc, une manière de distinguer ces changements de taille des effets de la structure serait de vérifier que la population étudié est plus proche d'un modèle panmictique que d'un modèle structuré. Nous avons montré qu'il est possible de distinguer deux modèles très simples dont un considérant une population panmictique avec un changement de taille et l'autre étant le *n-island model*. Pour ce faire nous avons proposé un test basé sur des valeurs de temps de coalescence de deux gènes (T_2).

Nous montrons aussi que dans le cas général, les valeurs de T_2 ne permettent pas distinguer un modèle panmictique avec changements de taille non contraints d'un modèle structuré. La raison est que, pour un modèle structuré quelconque (et d'ailleurs, pour n'importe quel modèle), il est possible de construire une fonction λ correspondant à des changement de taille de population sous un modèle panmictique qui fait que la distribution de T_2 sous le modèle panmictique est identique à celle de T_2 sous le modèle structuré. Par ailleurs, nous montrons que, lorsqu'on étudie une population structurée en se basant sur l'hypothèse de panmixie, on peut trouver des changements de taille complètement décorrélés de vrais changements de taille de population, pouvant même interpréter des changements de flux de gènes comme étant des *bottlenecks*. Cela laisse voir que l'inférence d'évènements démographiques dans le passé des populations structurées peut s'avérer problématique si on utilise une méthode basée sur l'hypothèse de panmixie.

Afin d'étudier l'histoire démographique d'une population structurée nous avons développé un modèle à partir du coalescent structuré. Ce modèle (nommé NIMC pour *N-island Markov Chain*) est basée sur un processus de Markov en temps continu et permet d'incorporer des changements dans le flux de gènes ainsi que dans la taille de la population. Nous montrons que les changements de taille inférés sous ce modèle correspondent à des vrais changements de taille de la population même lorsqu'elle est structurée.

Ce travail s'est réalisé dans le cadre d'une collaboration étroite entre biologistes, et mathématiciens et a ouvert de nombreuses voies de recherche. D'autres études seront nécessaires pour établir des critères permettant de décider, dans un cas général, si une population est plus proche d'un modèle panmictique ou bien d'un modèle de population structurée. Nous discutons dans le dernier chapitre comment le NIMC peut être utilisé pour obtenir la fonction de répartition de T_3, le temps d'apparition du premier ancêtre de deux gènes dans un échantillon de trois gènes. La densité jointe de T_2 et T_3 peut être utilisée pour reconstruire des arbres généalogiques en vue de donner des critères pour choisir le modèle le plus proche dans un cas réel.

J'ai par ailleurs co-dirigé plusieurs étudiants au cours de cette thèse. Je n'ai

pas présenté les résultats obtenus mais je peux noter que le travail du chapitre 2 qui étaient fondé sur les temps de coalescence a été étendu de deux manières. Nous avons testé la robustesse de l'estimation des paramètres du *n-island* en simulant de nombreux modèles plus réalistes, afin de voir quel types de biais pouvaient exister. Par ailleurs j'ai aussi développé cette approche en permettant de l'appliquer à des données de séquences et ai co-dirigé un étudiant qui a appliqué cette approche à des données réelles issues des espèces menacées pour lesquelles des données génomiques étaient accessibles.

Appendix A

PopSizeABC

PLOS | GENETICS

RESEARCH ARTICLE

Inferring Population Size History from Large Samples of Genome-Wide Molecular Data - An Approximate Bayesian Computation Approach

Simon Boitard[1,2]*, **Willy Rodríguez**[3], **Flora Jay**[4,5], **Stefano Mona**[1], **Frédéric Austerlitz**[4]

1 Institut de Systématique, Évolution, Biodiversité ISYEB - UMR 7205 - CNRS & MNHN & UPMC & EPHE, Ecole Pratique des Hautes Etudes, Sorbonne Universités, Paris, France, **2** GABI, INRA, AgroParisTech, Université Paris-Saclay, Jouy-en-Josas, France, **3** UMR CNRS 5219, Institut de Mathématiques de Toulouse, Université de Toulouse, Toulouse, France, **4** UMR 7206 Eco-anthropologie et Ethnobiologie, Muséum National d'Histoire Naturelle, CNRS, Université Paris Diderot, Paris, France, **5** LRI, Paris-Sud University, CNRS UMR 8623, Orsay, France

* simon.boitard@toulouse.inra.fr

Citation: Boitard S, Rodríguez W, Jay F, Mona S, Austerlitz F (2016) Inferring Population Size History from Large Samples of Genome-Wide Molecular Data - An Approximate Bayesian Computation Approach. PLoS Genet 12(3): e1005877. doi:10.1371/journal.pgen.1005877

Editor: Mark A Beaumont, University of Bristol, UNITED KINGDOM

Received: March 31, 2015

Accepted: January 27, 2016

Published: March 4, 2016

Data Availability Statement: Source code is available at https://forge-dga.jouy.inra.fr/projects/popsizeabc/. Cattle data was published in a previous study (Daetwyler et al, 2014, doi:10.1038/ng.3034). URLs to access the data are provided in this study.

Funding: SB was funded by the grant ANR-10-GENM-0014 from Agence Nationale de la Recherche (http://www.agence-nationale-recherche.fr) and the grant PEPS 2012 Bio-Maths-Info from Université de Toulouse (http://www.univ-toulouse.fr). FJ, SM and FA were funded by the grant ANR-12-BSV7-0012 from Agence Nationale de la Recherche

Abstract

Inferring the ancestral dynamics of effective population size is a long-standing question in population genetics, which can now be tackled much more accurately thanks to the massive genomic data available in many species. Several promising methods that take advantage of whole-genome sequences have been recently developed in this context. However, they can only be applied to rather small samples, which limits their ability to estimate recent population size history. Besides, they can be very sensitive to sequencing or phasing errors. Here we introduce a new approximate Bayesian computation approach named PopSizeABC that allows estimating the evolution of the effective population size through time, using a large sample of complete genomes. This sample is summarized using the folded allele frequency spectrum and the average zygotic linkage disequilibrium at different bins of physical distance, two classes of statistics that are widely used in population genetics and can be easily computed from unphased and unpolarized SNP data. Our approach provides accurate estimations of past population sizes, from the very first generations before present back to the expected time to the most recent common ancestor of the sample, as shown by simulations under a wide range of demographic scenarios. When applied to samples of 15 or 25 complete genomes in four cattle breeds (Angus, Fleckvieh, Holstein and Jersey), PopSizeABC revealed a series of population declines, related to historical events such as domestication or modern breed creation. We further highlight that our approach is robust to sequencing errors, provided summary statistics are computed from SNPs with common alleles.

(http://www.agence-nationale-recherche.fr). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Author Summary

Molecular data sampled from extant individuals contains considerable information about their demographic history. In particular, one classical question in population genetics is to reconstruct past population size changes from such data. Relating these changes to various climatic, geological or anthropogenic events allows characterizing the main factors driving genetic diversity and can have major outcomes for conservation. Until recently, mostly very simple histories, including one or two population size changes, could be estimated from genetic data. This has changed with the sequencing of entire genomes in many species, and several methods allow now inferring complex histories consisting of several tens of population size changes. However, analyzing entire genomes, while accounting for recombination, remains a statistical and numerical challenge. These methods, therefore, can only be applied to small samples with a few diploid genomes. We overcome this limitation by using an approximate estimation approach, where observed genomes are summarized using a small number of statistics related to allele frequencies and linkage disequilibrium. In contrast to previous approaches, we show that our method allows us to reconstruct also the most recent part (the last 100 generations) of the population size history. As an illustration, we apply it to large samples of whole-genome sequences in four cattle breeds.

Introduction

Reconstructing the ancestral dynamics of effective population size is important in several contexts. From a long term evolutionary perspective, the history of population size changes can be related to various climatic or geological events, and reconstructing this history allows studying the impact of such events on natural species [1]. This demographic history also provides a statistical null model of neutral evolution that can subsequently be used for detecting loci under selection [2, 3]. In conservation biology, the recent dynamics of effective population size in endangered species, as reconstructed from genetic data, can efficiently be used to decipher the time frame of a population decline, hence allowing to separate anthropogenic from natural factors [4].

Until recently, methods allowing to infer the history of population size changes from genetic data were designed for data sets consisting of a limited number of independent markers or non recombining DNA sequences [5–8]. However, the spectacular progress of genotyping and sequencing technologies during the last decade has enabled the production of high density genome-wide data in many species. New statistical methods accounting for recombination and scalable to the analysis of whole genome sequences are thus needed, in order to take advantage of this very rich source of information.

In this context, several promising approaches allowing to infer complex histories, including several tens of stepwise population size changes, have recently been proposed [9–13]. Some of them, called PSMC [9], MSMC [10] and diCal [11], are based on the Sequentially Markovian Coalescent (SMC or SMC') models [14, 15], an approximation of the classical coalescent with recombination [16], where coalescent trees are assumed to be Markovian along the genome. Thanks to this Markovian assumption, maximum likelihood estimates of past population sizes can be efficiently obtained from the observation of one (for PSMC) or several (for MSMC and diCal) diploid genomes. Another approach [12] is based on the length of Identity By State (IBS) segments shared between two chromosomes along the genome. Using an iterative search,

it aims at finding a history of past population size changes for which the expected distribution of IBS segment lengths matches that observed in one diploid genome.

While the above methods take advantage of whole-genome data, they are so far restricted to the analysis of small sample sizes. In the case of SMC based methods, this implies a limited resolution for the estimation of recent population sizes. Indeed, the most recent time at which these methods can infer population size is determined by the time to the most recent coalescence event occurring in the sample, which is older for small samples. For instance in humans, PSMC cannot estimate population sizes more recently than 400 generations (10,000 years) before present (BP), and MSMC cannot estimate these sizes more recently than 40 generations (1,000 years) BP. The most recent time for which an inference is possible will differ between species. In populations with small recent population sizes, coalescence events will occur at a higher rate than in larger populations, so the inference of recent history will be more accurate. Inference approaches based on the distribution of IBS segment length may be less affected by the use of small samples. Using this approach, estimations of population size in the Holstein cattle breed were obtained from a single genome even for the first few generations BP [12], and were in good agreement with estimations obtained from pedigree information in this breed [17–19]. However, the accuracy of the IBS approach used in this study has not been formally validated using simulations.

Another concern of the above methods is their sensitivity to sequencing errors. False positive SNPs can lead to a strong overestimation of population sizes in the recent past, i.e. in the first few hundred generations BP, both with PSMC [9, 12] and with the distribution of IBS segment length [12]. In contrast, false negative SNPs lead to underestimate population size at all time scales, but the magnitude of this effect is much weaker [12]. Efficient strategies for estimating these error rates and correcting the data accordingly have been proposed in [12]. However, the estimation step typically requires other sources of information than the sampled sequences, such as independent SNP chip data for the same individuals, which in many cases are not available. Phasing errors may also be an issue when inference is based on phased haplotype data, which is typically the case for MSMC [10] or diCal [11]. MSMC inference can also be based on unphased data, but this reduces the estimation accuracy [10].

Here we introduce a new statistical method named PopSizeABC, allowing estimating population size history from a sample of whole-genome sequences. One of the main motivations for developing this method is to take advantage of large sample sizes in order to reconstruct the recent history as well. Since statistical approaches based on the full likelihood of such samples seem currently out of reach, even with approximated models such as the SMC, we followed an Approximate Bayesian Computation (ABC) [20] approach, which simplifies the problem in two ways. First, this approach does not focus on the full likelihood of sampled genomes, but on the likelihood of a small set of summary statistics computed from this sample. Second, population size histories that are consistent with these observed summary statistics are inferred by intensive simulations rather than by complex (and generally intractable) mathematical derivations.

ABC is a popular approach in population genetics, which has already been applied to the analysis of large-scale population genetic data sets [21–25]. However, none of these previous studies tried to estimate complex population size histories involving a large number of population size changes. To address this question, we considered two classes of summary statistics: the folded allele frequency spectrum (AFS) and the average linkage disequilibrium (LD) at different physical distances. These two classes of statistics are very informative about past population size, and each of them is the basis of several inference approaches in population genetics [13, 26–31]. Therefore, combining them within an ABC framework seems very promising.

Applying our ABC approach to samples of 25 diploid genomes, simulated under a large number of random population size histories, we show that it provides, on average, accurate

estimations of population sizes from the first few generations BP back to the expected time to the most recent common ancestor (TMRCA) of the sample. This result is confirmed by the study of several specific demographic scenarios, where our method is generally able to reconstruct the population size history from present time back to the expected TMRCA, while PSMC or MSMC reconstruct it only for a limited time window.

We then apply this method to samples of 15 or 25 genomes in four different cattle breeds, which reveals interesting aspects of cattle history, from domestication to modern breed creation. Through this application to a real data set, we also illustrate how sequencing and phasing errors, if not taken into account, can have a dramatic influence on the estimated past population sizes. Our method is actually insensitive to phasing errors, because it uses unphased data. In addition, we show that a simple modification in the choice of summary statistics makes it robust to sequencing errors.

Results

Overview of the approximate Bayesian computation (ABC) estimation procedure

Following several recent studies [9–12], we modeled population size history as a stepwise constant process with a fixed number of time windows, where population size was constant within each window but was allowed to change from one window to the next. Time windows were defined in generations, for instance the most recent window went from one to ten generations before present (BP), and the most ancient window started 130,000 generations BP. This model allows approximating all simple demographic scenarios generally considered in population genetics studies (constant size, linear or exponential growth or decline, bottleneck . . .), as well as a large range of more complex demographic scenarios, provided population size changes occurred more recently than 130,000 generations BP.

Our estimation procedure was based on the observation of n diploid genomes sampled from the same population. We summarized this data set using two classes of summary statistics: (i) the folded allele frequency spectrum (AFS) of the sample, which includes the overall proportion of polymorphic sites in the genome and the relative proportion of those polymorphic sites with i copies of the minor allele, for all values of i between 1 and n, and (ii) the average linkage disequilibrium (LD) for 18 bins of physical distance between SNPs, from approximately 500 bp to 1.5Mb. We generated a very large number of population size histories, by drawing the population size in each time window from a prior distribution. For each history, we simulated a sample of n diploid genomes and computed a distance between the summary statistics obtained from this simulated sample and those obtained from the observed sample. A given proportion (called tolerance) of the most likely histories was accepted based on this distance. Finally, the joint posterior distribution of population sizes was estimated from the population sizes of accepted histories. Different statistical approaches were compared for this last estimation step.

A detailed description of the model and of the ABC procedure described above is provided in the Methods.

Accuracy of ABC estimation and relative importance of summary statistics

In order to optimize our ABC estimation procedure and to evaluate its average performance, we first applied it to a large number of genomic samples simulated under random population size histories. These pseudo-observed datasets (PODs) included 25 diploid genomes and 100 independent 2Mb-long regions. For each POD, population sizes were estimated by ABC, using

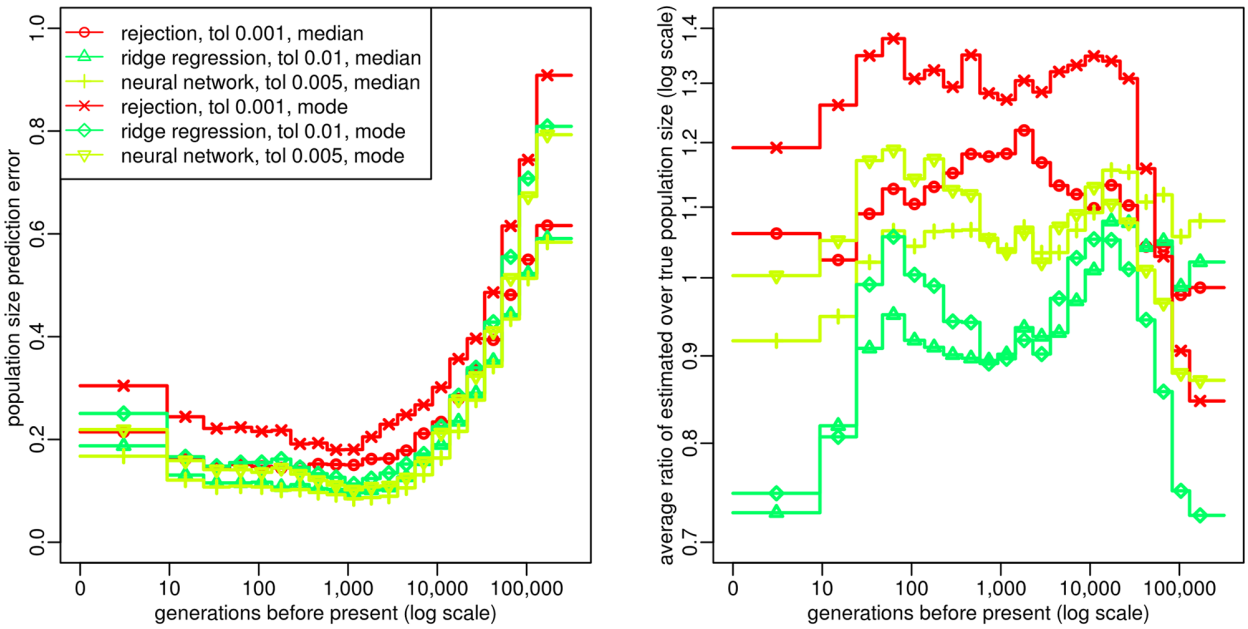


Fig 1. Optimization of ABC procedure. Prediction error (left panel) and bias (right panel) for the estimated population size in each time window, evaluated from 2,000 random population size histories (see Methods). Summary statistics considered in the ABC analysis were (i) the AFS and (ii) the average zygotic LD for several distance bins. These statistics were computed from $n = 25$ diploid individuals, using all SNPs for AFS statistics and SNPs with a MAF above 20% for LD statistics. The posterior distribution of each parameter was obtained by rejection, ridge regression [33] or neural network regression [32]. The tolerance rate used for each of these approaches was the one providing the lowest prediction errors, for different values from 0.001 to 0.05. Population size point estimates were obtained from the median or the mode of the posterior distribution. The prediction errors were scaled in order that point estimates obtained from the prior distribution would result in a prediction error of 1.

doi:10.1371/journal.pgen.1005877.g001

450,000 simulated datasets of the same size. These estimated values were compared with their true values for different tolerance rates and different ABC adjustment approaches to process the accepted histories. We found that the best procedure was to accept simulated histories with a tolerance rate of 0.005, to adjust their parameter values using a non linear neural network regression [32], and to summarize the resulting posterior distribution by its median. Indeed, point estimations of population sizes obtained by this procedure showed very small bias and the lowest prediction errors (PE) (Fig 1). Moreover, the posterior distributions of population sizes in each time window were correctly estimated, as shown by the accuracy of the 90% credible interval (S1 Fig, left), while the size of this credible interval was much lower than that obtained by the other adjustment approaches considered (S1 Fig, right). We used this procedure throughout the remaining of this study.

ABC provided accurate estimations of population sizes for a large range of times in the past (Fig 1). The best results were obtained from 10 to 5,000 generations BP, where the prediction error was below 0.1: this means that the average distance between true and estimated population sizes for this period of time was more than 10 times smaller than if the population sizes were estimated from the prior distribution. In the very recent past (from 0 to 10 generations BP), this prediction error was slightly larger but remained below 0.2. The prediction error also increased for times more ancient than 5,000 generations BP, while remaining quite low ($PE \leq 0.3$) until approximately 20,000 generations BP. This increase in prediction error above 5,000 generations BP can be related to a coalescence argument. At this time, the observed samples have coalesced to their common ancestor at most of the genomic regions, so the influence of demography on the current sample is reduced. Indeed, when rescaling time from

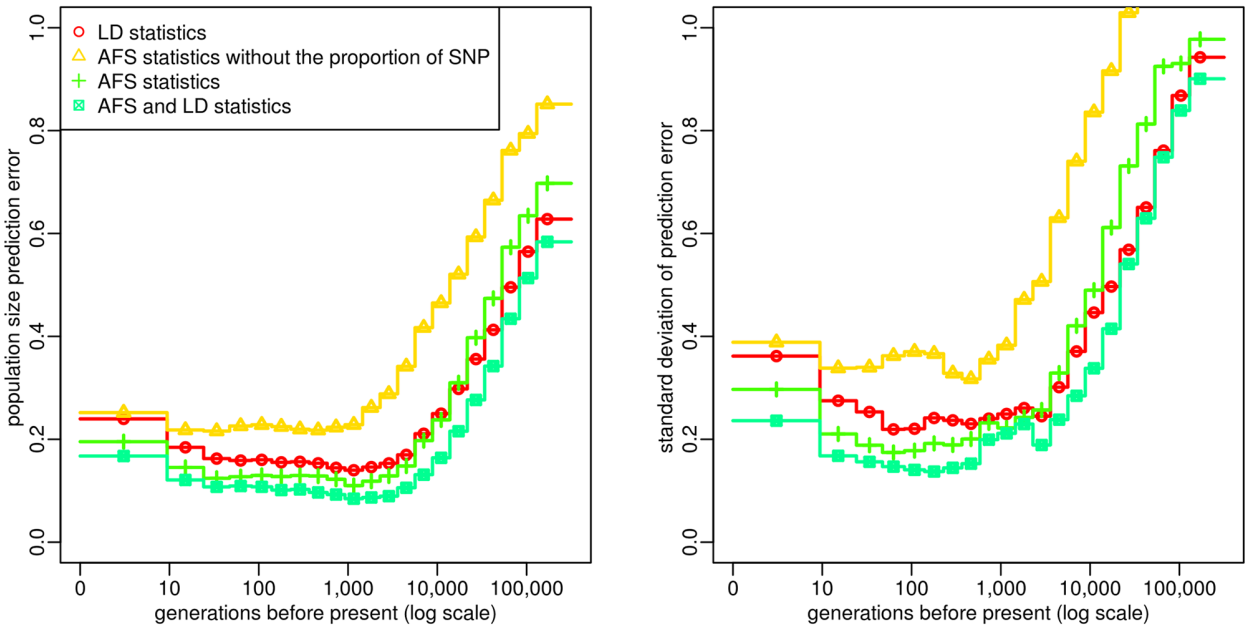


Fig 2. Accuracy of ABC estimation and relative importance of the summary statistics. Prediction error for the estimated population size in each time window (left) and standard deviation of this error (right), evaluated from 2,000 random population size histories. Summary statistics considered in the ABC analysis included different combinations of (i) the AFS (possibly without the overall proportion of SNPs) and (ii) the average zygotic LD for several distance bins. These statistics were computed from $n = 25$ diploid individuals, using all SNPs for AFS statistics and only those with a MAF above 20% for LD statistics. The posterior distribution of each parameter was obtained by neural network regression [32], with a tolerance rate of 0.005. Population size point estimates correspond to the median of the posterior distribution. The prediction errors were scaled in order that point estimates obtained from the prior distribution would result in a prediction error of 1.

doi:10.1371/journal.pgen.1005877.g002

generations to coalescent units (as described in Methods), we observed that the prediction error averaged over PODs started to increase shortly after the expected TMRCA (S2 Fig).

Our simulation study also highlighted the contribution of the different summary statistics. First, we found that population size history can be estimated quite well using either the AFS statistics alone or the LD statistics alone, but that combining the two classes of statistics clearly leads to the lowest PE for all time windows (Fig 2, left). As some demographic histories were more difficult to estimate than others, the PE differed between histories, but we observed that combining AFS and LD statistics allowed to reduce these differences (Fig 2, right). It also led to a reduction of the width of the 90% credible interval, as compared with the interval obtained using either class of statistics alone (S3 Fig). Another important advantage of combining AFS and LD statistics is to enable estimating the per site recombination rate. Indeed, the PE of this parameter was equal to 0.2 when using all statistics, versus 0.96 and 0.75 when using, respectively, AFS or LD statistics alone. Second, we found that using the polymorphic site AFS, i.e. the AFS without the overall proportion of SNPs, resulted in much higher PEs than using the full AFS (Fig 2). Third, we observed that computing LD at each SNP pair as a correlation between two vectors of n genotypes or as a correlation between two vectors of $2n$ alleles was equivalent in terms of PE (S4 Fig). This result implies that, with our approach, using unphased data rather than phased data will not decrease the estimation accuracy. Besides, computing LD from SNPs with relatively frequent alleles (MAF $\geq$ 5–20 %) resulted in lower PEs than computing it from all SNPs (S4 Fig). In the following, LD statistics were always computed from genotype data at SNPs with a MAF above 20%, unless otherwise specified.

Influence of the amount of data on ABC estimation

Another important question was to assess the amount of data that needs to be simulated and observed in order to achieve optimal accuracy. We first studied the influence of the number of simulated samples and found that increasing this number above 450,000 would not improve the estimation. Indeed, equally low PE and equally small (and accurate) confidence intervals could be obtained using 200,000 simulated samples (S5 Fig).

We then considered the influence of the genome length of observed and simulated samples (S6 Fig). As expected, PEs and the width of credible intervals decreased when the genome length increased. However, only small differences were observed between the performances obtained with 50 and 100 2Mb-long segments, and generating simulated data sets with much more than 100 2Mb-long segments (the default setup considered here) would become very challenging from a computational point of view (see the Methods for more details). For the analysis of observed data sets with a genome length above 200Mb, we thus considered the alternative strategy consisting in comparing observed statistics computed from the full genome (which is computationally very easy) with simulated statistics computed from a subset of the genome. We may think about these simulated summary statistics as an approximation of the genome-wide simulated statistics. To evaluate this strategy, we assumed that the genome length was 100 2Mb-long segments in the observed sample and 10 2Mb-long segments in the simulated samples (S7 Fig). Credible intervals were only slightly improved compared to using a genome length of 10 2Mb-long segments in both simulated and observed datasets, but PEs and their variance between scenarios were reduced, especially for the most recent and the oldest time windows, reaching values almost as low as those obtained when using 100 2Mb-long segments in both simulated and observed datasets. This strategy was thus applied in the further sections of the manuscript, where simulated statistics used for ABC estimation were computed from genomes made of 100 2Mb-long segments, independently of the genome length in the observed data.

We also studied the influence of sample size on population size estimations (S8 Fig). Comparing several sample sizes from $n = 10$ to $n = 50$, we observed that using large samples resulted in a more accurate estimation of population sizes in the first 100 generations BP. For instance, in the most recent time window, PE was equal to 0.153 for $n = 50$ versus 0.212 for $n = 10$, and the 90% credible interval was narrower (ratio between upper and lower bound of 36 versus 74). These improvements resulted from the fact that low frequency alleles, which are better captured from large samples, are very informative about recent population history. In contrast, population sizes at times more ancient than 10,000 generations BP were more accurately estimated from small samples, although the magnitude of this effect was lower than for recent population sizes (PE of 0.66 for $n = 50$ versus 0.63 for $n = 10$ in the most ancient window). This is likely due to statistical overfitting: increasing the sample size leads to increasing the number of AFS statistics, so if these additional statistics are not sufficiently informative they may introduce some noise and reduce the prediction ability of the model.

Finally, we found that computing AFS statistics only from SNPs exceeding a given minor allele frequency (MAF) threshold (from 5 to 20%) resulted in larger PEs and confidence intervals, except for the most ancient population sizes (S9 Fig). Again, this comes from the fact that low frequency alleles are very informative about recent population history. However, as we discuss later, introducing a MAF threshold might be necessary for the analysis of real data sets, so it is interesting to note that even with a MAF threshold of 20% the PE was not much larger than with all SNPs (0.24 versus 0.17 in the worst case).

Estimation of specific demographic scenarios using ABC

To illustrate the performance of our ABC approach, we then considered six specific demographic scenarios: a constant population size of 500, a constant population size of 50,000, a population size declining from 40,000 to 300 individuals between 3,600 and 100 generations BP, a population size increasing from 2,500 to 60,000 individuals between 1,500 and 250 generations BP, a population size experiencing one expansion from 6,000 to 60,000 individuals followed by a bottleneck of the same magnitude, between 34,000 and 900 generations BP, and a "zigzag" scenario similar to the previous one but including one additional bottleneck between 520 and 50 generations BP (see Fig 3 for more details). The decline scenario was chosen to mimic the estimated population size history in Holstein cattle [12], the expansion scenario was chosen to mimic the estimated population size history in CEU humans [10], and the "zigzag" scenario has been proposed in [10] as a typical example of very complex history. For each scenario, we simulated 20 PODs of 25 diploid genomes, each genome consisting in 500 independent 2Mb-long segments.

We observed that all PODs from a same scenario provided very similar ABC estimations (Fig 3). This suggests that increasing the observed genome length would not improve the obtained estimations, at least with the levels of mutation (1e-8 per bp) and recombination (5e-9 per bp) and the population sizes considered here. Besides, as expected from our previous simulation results, population size history could be reconstructed for all scenarios from a few generations BP back to at least the expected TMRCA of the sample, with the only two exceptions described below.

First, population size estimations in the most recent time window (less than 10 generations BP) often showed a slight bias towards intermediate values, as can be seen in the large constant size scenario, the decline scenario and the expansion scenario. This partly comes from the fact that we estimated population size by the median of the posterior distribution, which tends to shrink it away from our prior boundaries. When estimating population sizes from the mode of the posterior distribution, we were able to better reconstruct the very recent population size in these three scenarios (S10 Fig). Nevertheless, using the mode also brought other issues: it led to less smooth population size histories (S10 Fig) and, on average, to larger PEs than using the median (Fig 1). Second, the zigzag scenario was incompletely reconstructed: the initial increase of population size and the subsequent first bottleneck could be recovered, but the second bottleneck was replaced by a slow decline.

In order to explore why ABC failed to fully reconstruct this zigzag history, we considered five variants of this scenario (S11 Fig). For a zigzag scenario with smaller population sizes than the original one (ten times lower in all time windows), we observed that ABC could recover the full sequence of expansions and contractions (S11 Fig, top right). This was also the case when only one of the two bottlenecks of this "zigzag small" history was simulated (S11 Fig, bottom). In contrast, when only the most recent bottleneck of the "zigzag large" scenario was simulated, ABC could still not reconstruct it (S11 Fig, middle left). Actually, the decline wrongly estimated by ABC in this case led to very similar summary statistics as the true bottleneck (S12 Fig), and the population size trajectory corresponding to the true bottleneck was included in the 90% credible interval inferred by ABC (S11 Fig, middle left). We also observed that PODs simulated under the wrong decline history would lead to very similar ABC estimations that those simulated under the true bottleneck history (S11 Fig, middle, left vs right). These results suggest that the accuracy of our ABC approach is not strongly affected by the complexity (i.e. the number of expansions and declines) of the true history, but that some specific demographic events, in particular those implying recent population size changes in large populations, can be difficult to identify using this approach. This conclusion was supported by the study of four additional

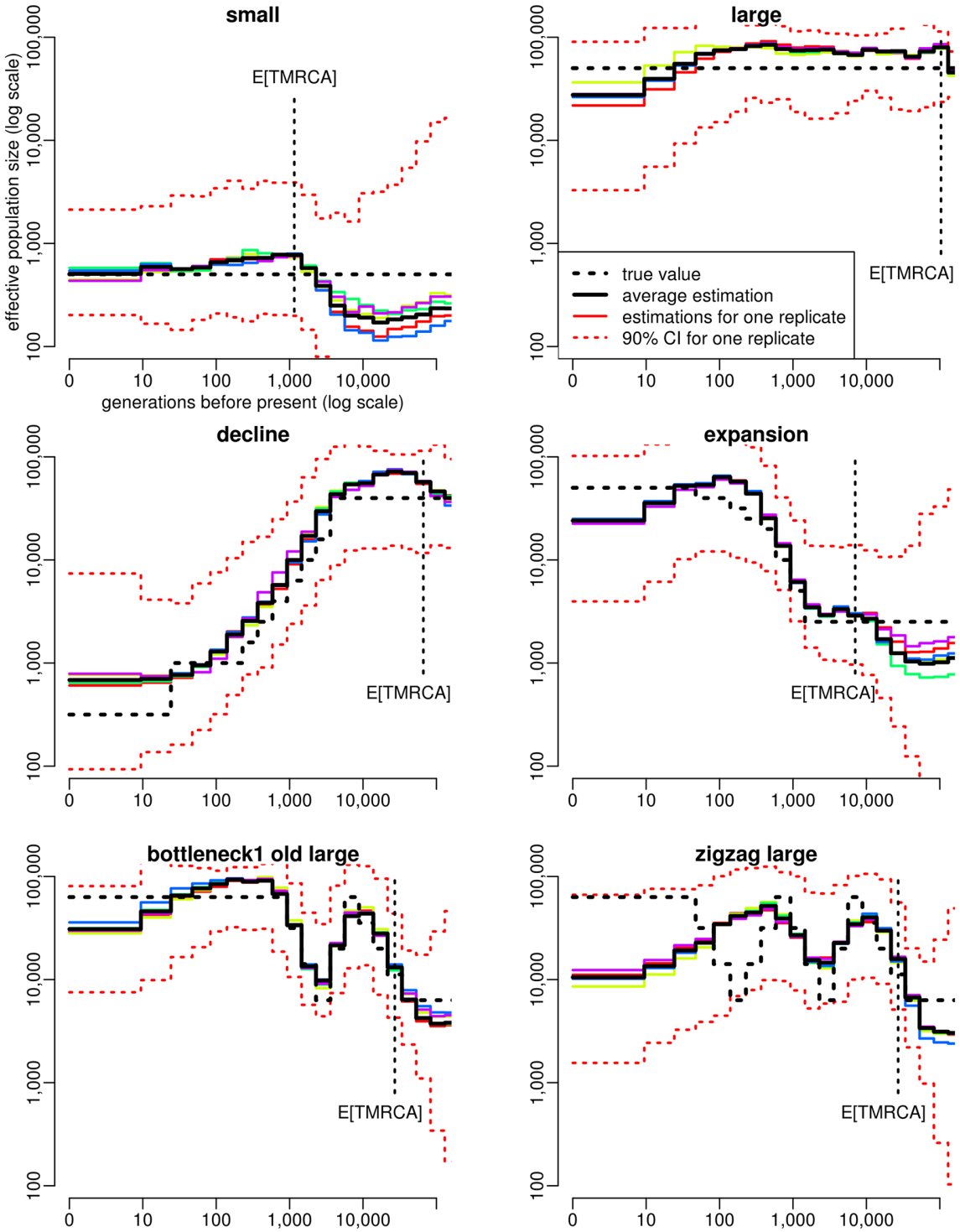


Fig 3. Estimation of population size history using ABC in six different simulated scenarios. a small constant population size ($N = 500$, top left), a large constant population size ($N = 50{,}000$, top right), a decline scenario mimicking the population size history in Holstein cattle (middle left), an expansion scenario mimicking the population size history in CEU human (middle right), a scenario with one expansion followed by one bottleneck (bottom left) and a zigzag scenario similar to that used in [10] (bottom right), with one expansion followed by two bottlenecks. For each scenario, the true population size history is shown by the dotted black line, the average estimated history over 20 PODs is shown by the solid black line, the estimated histories for five random PODs are shown by solid colored lines, and the 90% credible interval for one of these PODs is shown by the dotted red lines. The expected time to the most recent

common ancestor (TMRCA) of the sample, *E*[*TMRCA*], is indicated by the vertical dotted black line. Summary statistics considered in the ABC analysis were (i) the AFS and (ii) the average zygotic LD for several distance bins. These statistics were computed from $n = 25$ diploid individuals, using all SNPs for AFS statistics and SNPs with a MAF above 20% for LD statistics. The posterior distribution of each parameter was obtained by neural network regression [32], with a tolerance rate of 0.005. Population size point estimates were obtained from the median of the posterior distribution.

doi:10.1371/journal.pgen.1005877.g003

complex scenarios, implying similar expansions and declines as in S11 Fig but in a different order, i.e. the first event was a bottleneck and it was followed by a population decline (S13 Fig). Except the recent part of the "bottleneck2 recent large" scenario (S13 Fig, top left), all aspects of these histories occurring more recently than the expeted TMRCA were accurately reconstructed by ABC.

Because one of our objectives was to estimate the population size history in taurine cattle, we studied more precisely the continuous decline scenario that is expected in this species [12], and evaluated if variations from this scenario could be detected by ABC (S14 Fig). We found that a decline of the same magnitude (from 40,000 to 300), but occurring suddenly either 200 generations BP (top right) or 1,000 generations BP (middle left), would lead to a clearly distinct ABC estimation, although ABC had a tendancy to smooth population size changes. We also considered two scenarios where population size increased again after the sudden decline occuring 1,000 generations BP, either quickly to a relatively high value (5,000, middle right) or more recently to a lower value (1,000, bottom left). In the two scenarios, both the bottleneck phase and the recovery phase could be inferred by ABC. Finally, we studied an alternative scenario where the initial continuous decline was followed by a sudden decline to 100 between 230 and 140 generations BP and by a later recovery to 1,000 (bottom right). Assuming generation time in cattle is about 5 years, the time frame of this bottleneck (between 1,150 and 700 years BP) would correspond to the Middle Age period, where cattle population sizes may have decreased drastically because of wars, famines and cattle plagues [34]. Again, we found that ABC should be able to distinguish this scenario from a simple continuous decline.

Comparison with MSMC

For each scenario of Fig 3, we also analyzed five simulated samples with MSMC [10], using two, four or eight of the haplotypes from each sample. When applied to two haplotypes, MSMC is an improved version of PSMC [9], a software that has been used to estimate population size history in many different species within the last few years [35–38]. In our simulations, MSMC based on two haplotypes provided a very accurate estimation of the population size history within a time window starting between a few hundreds and a few thousands generations BP, depending on the scenario, and finishing after the expected TMRCA of 50 haplotypes (Fig 4). Within this time window, estimations obtained by MSMC from the five replicates were all very close to the true history, even more than ABC estimations. Outside this window however, population size histories estimated by MSMC often had a totally different trend than the true history, (see for instance the small constant size or the decline scenario), with large differences observed between samples (see for instance the expansion scenario). Similar results were obtained when using MSMC with four (S15 Fig) or eight (S16 Fig) haplotypes, except that the time window where accurate population sizes could be obtained was shifted towards recent past, as already shown in [10]. This comes from the fact that MSMC inference is based on the time to the most recent coalescence event, which decreases when the sample size increases. Thus, reconstructing the entire population size history is generally not possible from a single MSMC analysis. This would require concatenating different parts of the history estimated independently using different sample sizes, which might be quite challenging in a real data analysis, because the bounds to consider for such a concatenation are unknown. Besides, for four out of the six scenarios (the

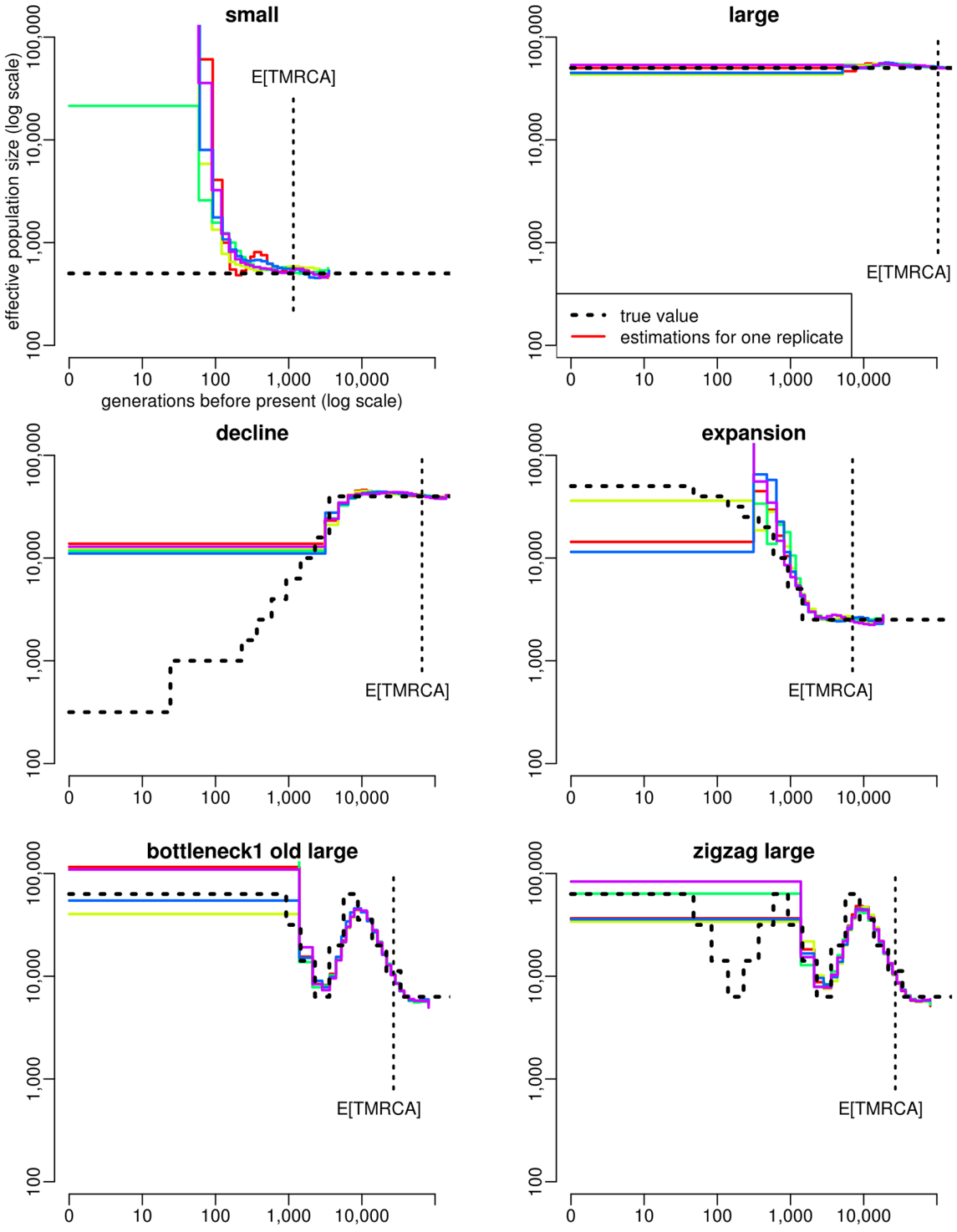


Fig 4. Estimation of population size history using MSMC with two haplotypes in five different simulated scenarios. For each scenario, the five PODs considered for MSMC estimation were the same as in Fig 3. The expected TMRCA shown here is also the same as in Fig 3, it corresponds to samples of 50 haploid sequences.

doi:10.1371/journal.pgen.1005877.g004

small constant size, the expansion, the bottleneck and the zigzag), population sizes at times more recent than approximately 100 generations BP could not be estimated by any MSMC analysis. Indeed, the analysis with eight haplotypes, which is expected to be the most accurate for reconstructing recent demography, provided unstable results for these scenarios. Several other cases where MSMC failed to reconstruct properly the recent history were observed among the additional scenarios tested with ABC, as for instance in the "bottleneck cattle middle age" scenario (S17 Fig and S18 Fig) for which the recent bottleneck was not detected.

Finally, it is important to note that the simulated data that we used in these MSMC analyzes were assumed to be perfectly phased. However, real data consist generally in statistically inferred haplotypes, which can typically include from 1 to 10 switch errors per Mb and individual, even when using recent phasing algorithms and large sample sizes [39]. In our simulations, analyzing phased data with such switch error rates often biased MSMC estimations, especially for the most recent part of the demographic history (S19 Fig). To avoid this issue, MSMC can in principle be run from unphased data, but we found that this would also affect the estimation accuracy (S19 Fig, right column).

Application to NGS samples in cattle

We applied our ABC approach to estimate the population size history in four cattle breeds, using large samples of diploid genomes recently published by the 1,000 bull genomes project [40]. An important issue when analyzing NGS data is the potential influence of sequencing and phasing errors on the estimations. To investigate this question, we first evaluated how these errors affect the summary statistics considered in our ABC approach. We considered a set of 12 Holstein animals for which the haplotypes inferred from NGS data within the 1,000 bull genomes project could be compared with those inferred from 800K SNP chip data obtained independently from another project. Assuming that 800K data are free of genotyping errors, we computed the summary statistics from these data and checked whether similar values could be obtained from NGS data at the same positions (S20 Fig). We found that the average gametic LD (i.e. the LD computed from haplotype data) was significantly smaller with NGS data than with 800K data at long physical distances, but not at short ones. This likely comes from an increased level of phasing errors in NGS data as compared to 800K data. Indeed, such errors tend to artificially break the correlation between SNPs within each individual, which reduces LD. Besides, as they are relatively rare, we expect their influence to be significant only when comparing SNPs at large physical distance.

In contrast, the average zygotic LD (i.e. the LD computed directly from genotype data) was identical for the NGS and the 800K data. We also observed a perfect match between the polymorphic site AFS obtained from the NGS data subsampled at 800K positions, and from 800K data. Finally, the overall proportion of SNPs was similar in the two types of data. More precisely, based on the 800K positions and the sample of 12 individuals, we found approximately 0.5% of false positive SNPs, i.e. positions that were found polymorphic when using NGS data but not when using 800K data (S21 Fig, left), and approximately 5% of false negative SNPs, i.e. positions that were found polymorphic when using 800K data but not when using NGS data (S21 Fig, right). Besides, the proportion of false negative SNPs did not depend on the true allele frequency (i.e. the allele frequency in the 800K data), so it should not distort the AFS. Overall, these results suggest that our summary statistics, when computed from genome wide unphased NGS data, should not be affected by sequencing and phasing errors. However, the above comparison does not really allow to evaluate the influence of false positive SNPs when analyzing genome wide NGS data, because the 800,000 positions of the SNP chip are strongly enriched in true SNPs compared to the three billions of positions of the entire genome.

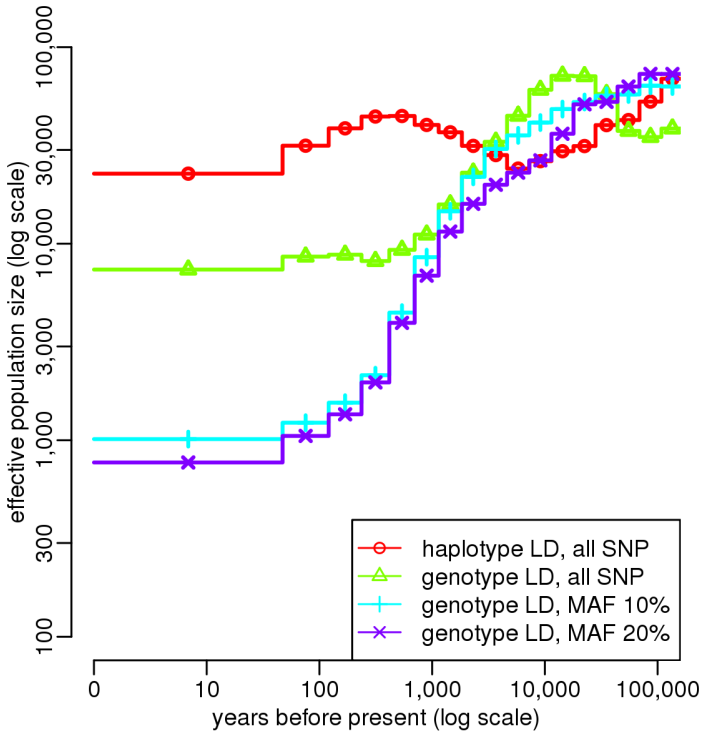


Fig 5. Influence of phasing and sequencing errors on ABC estimation. Estimation of population size history in the Holstein cattle breed using ABC, based on whole genome NGS data from $n = 25$ animals. Summary statistics considered in the ABC analysis were (i) the AFS and (ii) the average LD for several distance bins. LD statistics were computed either from haplotypes or from genotypes, using SNPs with a MAF above 20%. AFS statistics were computed using either all SNPs or SNPs with a MAF above 10 or 20%. The posterior distribution of each parameter was obtained by neural network regression [32], with a tolerance rate of 0.005. Population size point estimates were obtained from the median of the posterior distribution. Generation time was assumed to be five years.

doi:10.1371/journal.pgen.1005877.g005

To overcome this limitation, we studied directly the influence of sequencing and phasing errors on ABC estimations, by analyzing one sample of 25 Holstein genomes with slightly different combinations of summary statistics (Fig 5). When LD was computed from haplotypic data, the estimated recent population size was above 20,000 individuals, which seems quite unrealistic given that the estimated current effective size of this breed is generally of an order of 100 [17–19, 41]. This discrepancy likely resulted from the average LD at large physical distances, which was artificially reduced by phasing errors, as discussed above. Computing LD from genotypic data, we obtained more realistic results, with a recent population size of 7,000. However, there was a great difference between the estimation obtained when computing AFS statistics from all SNPs, and that obtained when computing these statistics only from SNPs with a MAF above 10% (Fig 5). Such a large difference was not expected from simulations, neither on average over multiple random histories (S9 Fig, middle) nor in the particular cases of a constant or declining population (Fig 3 vs S22 Fig). Thus, it must result from the influence of false positive SNPs, which are much more likely to produce low frequency alleles (S21 Fig, left). In contrast, there was little difference between the estimations obtained when computing AFS statistics with a MAF threshold of 10 or 20%, which strongly suggests that these strategies are both robust against sequencing errors, at least for this particular dataset. To be conservative, we used a MAF threshold of 20% for the final analysis of the four breeds.

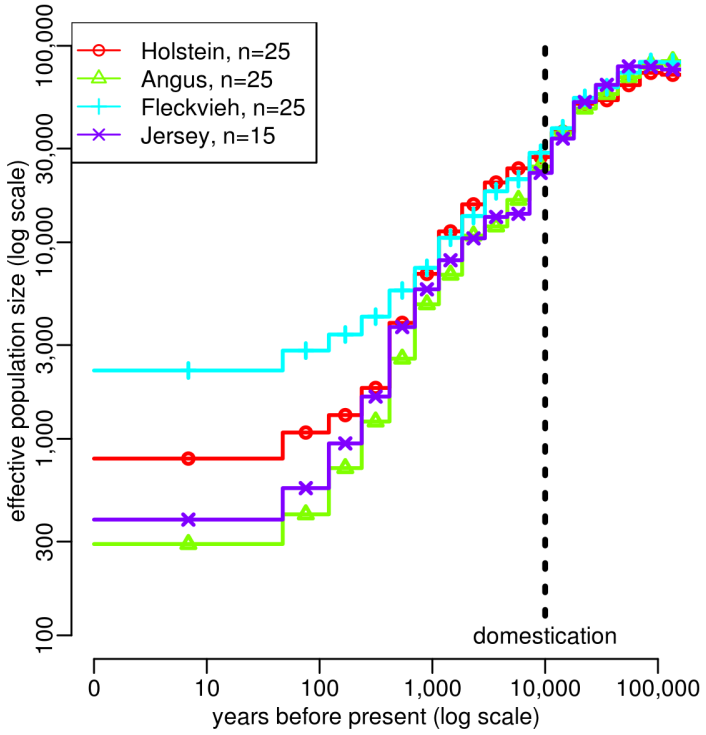


Fig 6. Estimation of population size history in four cattle breeds using ABC. Angus (n = 25 animals), Fleckvieh (n = 25), Holstein (n = 25) and Jersey (n = 15). Estimations were obtained independently in each breed, based on whole genome NGS data from sampled animals. Summary statistics considered in the ABC analysis were (i) the AFS and (ii) the average zygotic LD for several distance bins. These statistics were computed using SNPs with a MAF above 20%. Other parameter settings are the same as in Fig 5.

doi:10.1371/journal.pgen.1005877.g006

This analysis outlined several interesting features of cattle demographic history (Fig 6). Before 10,000 years BP, the population sizes estimated in the four breeds were very similar, in agreement with the fact that all four breeds descend from a same ancestral population, i.e. the initial *Bos taurus* population which resulted from the domestication of the wild aurochs, *Bos primigenius*, approximately 10,000 years BP [42]. This common estimated history is characterized by a population decline starting approximately 50,000 years BP. In particular, a sharper decrease was observed from approximately 20,000 years BP, which could correspond to the intensification of anthropogenic effects like hunting or later herding [42]. Shortly after domestication, the inferred population size histories could be divided into two groups, Holstein and Fleckvieh on one hand, Angus and Jersey on the other hand. This is consistent with the origin of these breeds: Holstein and Fleckvieh ancestors were brought into Europe through the *Danubian route* approximately between 7,500 and 6,000 years BP, while Angus and Jersey have more diverse origins and partly descend from animals that were brought into Europe through the *Mediterranean route* approximately between 9,000 and 7,300 years BP [43, 44]. Population size histories in the four breeds finally diverged during the last 500 years, which is consistent with the progressive divergence of these breeds induced by geographic isolation and, from the 18th century, by the creation of modern breeds [45]. This lead to recent effective population sizes of 290 in Angus, 390 in Jersey, 790 in Holstein and 2,220 in Fleckvieh.

The 90% credible intervals associated to these estimated population size histories are shown in S23 Fig. We performed posterior predictive checks by sampling population size histories from the posterior distributions and simulating new genomic samples from these histories

[46]. The summary statistics obtained from these samples were similar to those observed in the real data (S24 Fig). We also checked that the best simulated histories provided summary statistics that were indeed similar to the observed summary statistics (S25 Fig). Finally, we note that point estimations of the average per site per generation recombination rate were quite similar between breeds: it was equal to 3.66e-9 in Holstein, 3.89e-9 in Fleckvieh, 4.58e-9 in Jersey and 5.00e-9 in Angus.

Discussion

Methodological contribution

Applying our ABC approach to genomic samples simulated under a large number of random population size histories, we showed that it provides, on average, accurate estimations of population sizes from the first few generations BP back to the expected TMRCA of the sample. Because the estimation accuracy depends on the true population size history, we also analyzed genomic samples simulated under 20 specific demographic scenarios with various levels of complexity: a constant population size (2 scenarios), a monotonic decrease (3 scenarios) or expansion (1 scenario), a single bottleneck (3 scenarios), a single bottleneck plus an additional expansion or decrease (9 scenarios) or two bottlenecks plus an additional expansion (2 scenarios). For most of these scenarios, PopSizeABC could reconstruct the population size history from present time back to the expected TMRCA of the sample. Within this time limits, the only situations where the ABC point estimates were very different from the true history were those implying a decline or expansion occurring in a large population (more than 5,000 individuals) within the last few hundreds generations. Indeed, when large population sizes are combined with frequent population size changes (in our model, recent time windows are also the shortest ones), each time window represents a very small part of the coalescent history, which explains why these scenarios are particularly difficult to reconstruct. However, in these situations, the true history was still included within the 90% credible interval, and the increased width of this interval compared to other time windows suggested that the point estimate was less reliable. Similarly, in all scenarios, the width of the credible interval increased rapidly for times that were more ancient than the expected TMRCA, which corresponds thus to the upper bound of the time period where ABC estimation could be trusted.

Interestingly, we observed that PopSizeABC behaved quite differently from MSMC [10], a recent full-likelihood SMC-based method allowing to analyze multiple diploid genomes. On one hand, for the 20 scenarios considered here, MSMC estimated more accurately than PopSizeABC the population sizes at several time points. This was expected because ABC inference implies a much larger degree of approximation than MSMC inference. On the other hand, the total time period for which each demographic history could be correctly reconstructed with a single MSMC analysis was much smaller than with ABC. Besides, in most scenarios, recent population sizes (in the first 100 generations BP or even more) could not be inferred by any MSMC analysis, while they could be inferred by ABC. In our study of cattle demography, reconstructing the population size history for this recent period allowed to highlight the specificity of each breed. In many other situations, and especially in a conservation perspective, estimating recent demography is actually crucial.

The better performance of ABC to reconstruct recent population size history is partly explained by the possibility of using larger samples. We generally considered samples of 25 diploid genomes, which resulted in more accurate estimations of population sizes in the last 30 generations than using only 10 diploid genomes (S8 Fig). Indeed, large samples contain rare alleles. Since these alleles result from mutations that occurred in the most recent part of the coalescent tree, their relative proportion in the AFS is informative about the recent variations

of population size. Interestingly, gaining accuracy for recent time periods by increasing the sample size had no strong negative impact on the reconstruction of the older demographic history (except for times older than the TMRCA), contrary to what was observed with MSMC. The use of LD statistics must also contribute to the reconstruction of recent demography because, in our simulations, predictions of population sizes at times more recent than 100 generations BP were still acurate when rare alleles were removed (S9 Fig). As discussed below, the average LD at long physical distances is expected to reflect the recent population size [26].

Following previous studies [26, 30, 47], we used in our ABC approach the average LD over different bins of physical distance in order to get information about population sizes at different times in the past. In a finite population, LD results from a balance between drift and recombination. This implies that LD between markers at long recombination distance mostly reflects recent population sizes, while LD at short recombination distance also reflects ancient population sizes [48]. To illustrate this, we computed our LD statistics for several simulation scenarios consisting in a sudden expansion with fixed magnitude but occurring at different times in the past (S26 Fig, left). As expected, we observed that LD statistics at long distance were similar to those of a large population, thus reflecting the recent population size, while LD statistics at small distance were similar to those of a small population, thus reflecting the ancient population size. Besides, the more recent the expansion, the larger the distance required to observe a LD level reflecting the large (recent) population size. Similarly, for decline scenarios, markers at long (resp. short) distance were most of the time found to reflect the LD level in a small (resp. large) population (S26 Fig, right; see the legend for more details)).

This relation between the recombination distance and the time horizon can even be described more precisely. If population size is assumed to change linearly over time, it can be shown that the expected r^2 between SNPs at recombination distance c is approximately equal to

$$E[r^2] \approx \frac{1}{a + 4Nc} \quad (1)$$

where N is the effective population size at time $1/(2c)$ BP and a is a constant depending on the mutation model [26]. The evolution of population size through time can thus be reconstructed by computing the average r^2 for different bins of recombination distance, and then inverting the formula in Eq (1) [26, 47]. However, several authors pointed out that this approximation is unsatisfactory, especially for non constant demography [49, 50], and could lead to wrong estimations of past population sizes [50, 51]. Our ABC approach overcomes this issue, because r^2 values estimated from the data are not compared to approximate theoretical predictions, but to simulated r^2 values. Using this approach, we could demonstrate that these statistics contain useful information about the population size history (Fig 2). We further demonstrated two important properties of LD statistics in the context of population size inference (S4 Fig). First, computing r^2 from genotypes is as informative as computing it from haplotypes, in the sense that it leads to similar PEs. Second, removing rare SNPs (at least those with MAF below 5%) when computing this LD measure reduces PE.

In our simulations, ABC inferences based on AFS statistics alone also provided accurate estimations of population sizes at different times in the past (Fig 2). Theoretical studies have demonstrated that complex population size histories can be estimated from AFS statistics [52], and these statistics are already the basis of several inferential approaches in population genetics [13, 27–29, 31]. In particular, two recent studies implemented composite-likelihood approaches to estimate population size through time in a single population [13, 31], and obtained convincing results on simulated data. We do not expect that our ABC approach based on AFS statistics alone would improve the point estimations obtained by these approaches, and analyzing very large samples (i.e. hundreds or thousands of individuals) would certainly be

much more challenging with ABC due to the simulation step. However, one advantage of ABC is to provide credible intervals, which allow to quantify the degree of confidence associated to a given point estimation.

Moreover, one important conclusion of our work is that combining AFS and LD clearly improves, on average, the estimation of population sizes (Fig 2). This stems from the fact that these two classes of statistics are not informative for the same demographic scenarios. While prediction errors obtained from AFS or LD statistics were quite similar for scenarios with little population size variations (S27 Fig, top panels), better predictions were obtained from AFS (resp. LD) statistics when the main trend of the population size history was an expansion (resp. a decline) (S27 Fig, bottom panels). These differences were mainly due to the predictions obtained from AFS statistics, which were much better for expansion scenarios than for decline scenarios. Indeed, population declines accelerate the rate of recent coalescence events compared to old ones. Combined with the fact that the time intervals between recent coalescence events are intrinsically shorter than between old ones (because coalescence rates are proportional to the square of the sample size), this tends to produce coalescence trees where only the few oldest branches have a substantial length. In other words, the recent topology of coalescence trees in decline scenarios is very hard to infer based on observed data, making it difficult to estimate population size variations from the AFS. These results are consistent with those from a recent study [53], which showed that, for the inference of single bottleneck events, including some linkage information was more efficient than using the AFS alone. Actually, one interesting conclusion of S27 Fig is that combining LD and AFS statistics always improves the prediction compared to using either one or the other class of statistics alone, whatever the family of scenarios we considered.

This conclusion was also supported by the study of several specific scenarios: some could be accurately reconstructed from AFS statistics alone but not from LD statistics alone (Fig 7, top), and vice versa (Fig 7, middle), but the prediction obtained when combining AFS and LD statistics was always close to the best of the two. In other scenarios, neither AFS or LD statistics alone allowed to correctly estimate the demographic history, and using them jointly was therefore essential (Fig 7, bottom). Finally, in many scenarios, ABC estimation based either on LD statistics alone or AFS statistics alone performed already very well, but the advantage of combining AFS and LD statistics clearly appeared when using a MAF threshold that reduced the information brought by AFS statistics (S28 Fig). Besides these effects on population size estimation, note that combining AFS and LD statistics allowed to estimate the average per site recombination rate, which was not possible using either one or the other class of statistics alone.

The genome wide distribution of the length of IBS segments shared between two chromosomes could provide another interesting class of summary statistics for ABC, because several recent studies showed that it is very informative about population demography [12, 54]. However, we found that applying ABC from a set of statistics related to this distribution, rather than from AFS and LD statistics, resulted in larger PEs of population sizes more recent than 100 generations BP (S29 Fig). This is likely due to the much smaller number of individuals simultaneously considered in IBS statistics. When IBS statistics were used in addition to AFS and LD statistics, no significant improvement was observed compared to the combination of AFS and LD statistics. Besides, the estimation of recent population demography is mainly influenced by the frequency of long IBS segments, which might be difficult to estimate in practice due to sequencing errors [12, 54]. Thus, we did not further investigate the inclusion of these statistics in our approach.

Several previous studies implemented ABC approaches based on genome-wide data to infer population genetics models [21–25]. However, none of these studies focused on the estimation of population size through time using complex step-wise models, as we did here. In a Bayesian

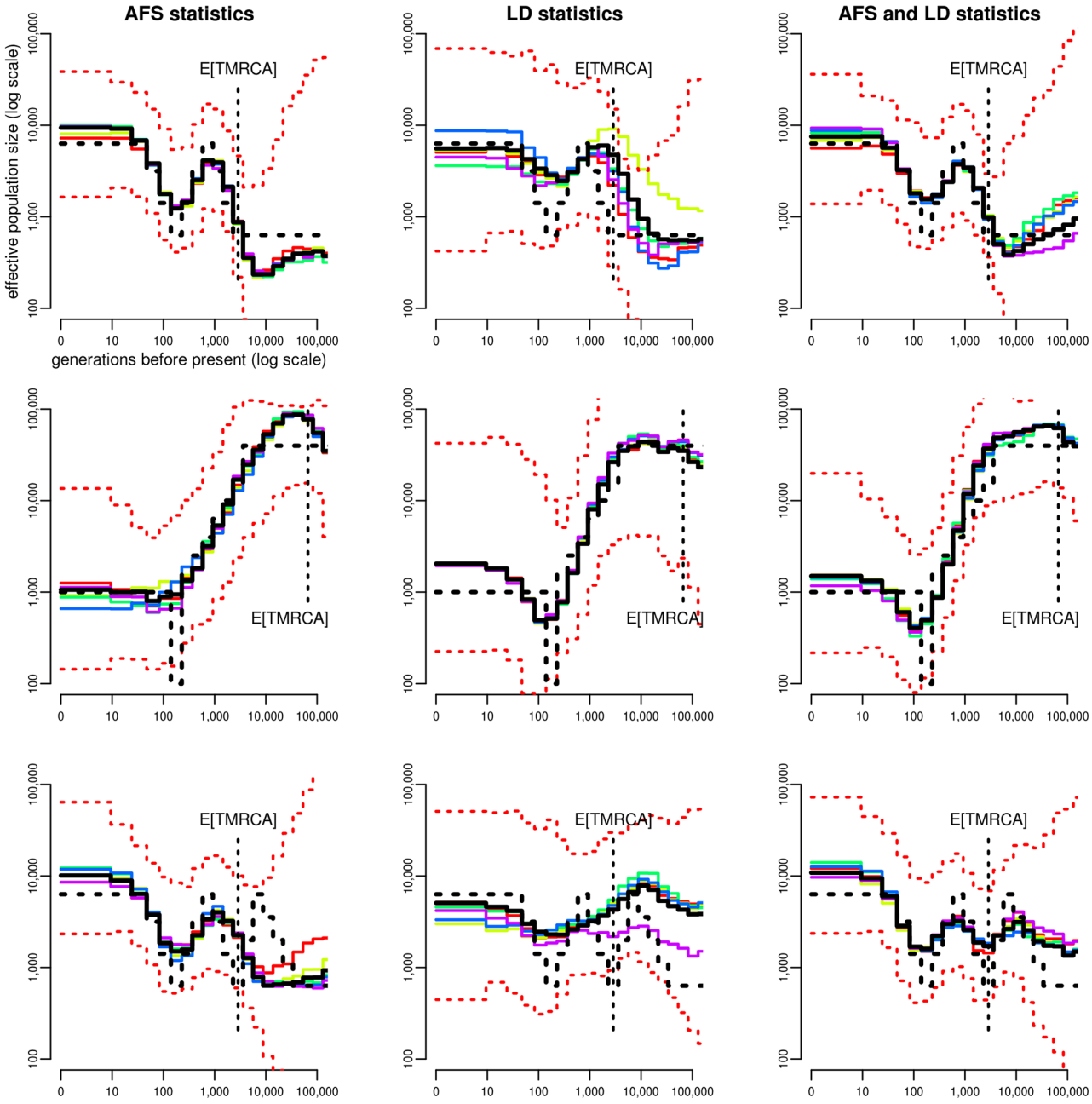


Fig 7. Comparison of summary statistics for the estimation of population size history in three scenarios. "bottleneck1 recent small" (top), "bottleneck cattle middle age" (middle) and "zigzag small" (bottom). Summary statistics considered in the ABC analysis were either the AFS statistics alone (left column), the LD statistics alone (middle column), or the AFS and LD statistics together (right column). All other settings are similar to Fig 3, as well as the legend.

doi:10.1371/journal.pgen.1005877.g007

perspective, this specific question had, so far, only been adressed using a small number of independent non-recombining loci [5–8]. Another originality of our study is to use LD summary statistics that can only be computed from relatively long DNA sequences (at least 2Mb) with recombination, while almost all previous genome-wide ABC studies (but see [23]) considered short loci ($\leq$ 20kb long). Even with modern computer facilities, simulating hundreds of thousands of long DNA sequences required some optimization adjustments. One of them was to

reduce the space of possible simulated histories to the most realistic ones by setting constraints on the prior distributions of population sizes (see Methods). Another one was to allow simulated and observed samples to differ in two different ways. First, the total genome length was generally smaller in simulated samples than in the observed sample, which resulted in lower prediction errors than reducing the genome length in the observed sample down to the one that could be efficiently achieved in simulated samples (S7 Fig). Second, when analyzing the cattle data, the simulated summary statistics were computed from independent 2Mb-long segments, although the observed ones were computed from contiguous 2Mb-long segments. Indeed, simulating data under the coalescent with recombination becomes extremely difficult for long sequences. This second approximation cannot bias the estimations, because the correlation structure between segments has no impact on the expected value of summary statistics. Similar to the genome length, the correlation structure of the genome only affects the precision (i.e. the estimation variance) of summary statistics. Despite of the additional correlation, computing summary statistics in cattle using the entire genome ($\approx$ 1,250 contiguous 2Mb-long segments) likely resulted in a higher precision, and thus in a more acurate estimation, than using a subset of 100 independent 2Mb-long segments.

Analyzing real data sets with our approach presents several important advantages. First, our approach is designed to be applied to totally unphased data. Indeed, AFS statistics are deduced from the allele frequencies at all SNPs, which can be computed directly from genotypes. LD statistics are also computed from genotypes, although the common practice in population genetics is to compute them from haplotypes. LD statistics computed from genotypes are not identical to LD statistics computed from haplotypes, but they lead to similar estimations of population sizes. As observed in the analysis of the cattle data (Fig 5), phasing errors can have dramatic effects on the estimated histories, and they would certainly affect the inference for all populations where the experimental design prevents from phasing the data with high accuracy. Moreover, the SNP data handled by our method can be unpolarized, i.e. it is not necessary to know which of the alleles at a given SNP is ancestral. Using polarized data would probably improve the estimations, as this would allow computing the unfolded rather than folded AFS. However, inferring ancestral alleles is not always possible and is prone to errors, so we chose to focus on statistics computable for all datasets. Finally, based on the analysis of NGS data in cattle, we showed that our approach can easily be made robust to sequencing errors by computing summary statistics only from SNPs with common alleles (MAF $\geq$ 10 or 20%, Fig 5). This modification is expected to increase the population size prediction errors and the width of credible intervals if the dataset contains no sequencing errors (S9 Fig), but this seems by far preferable to the large biases caused by sequencing errors, as illustrated by our study and several previous ones [9, 12].

One consequence of sequencing errors is to create wrong SNP calls in the data, at genomic positions where the observed sample is actually not polymorphic. Because these wrong SNPs are generally associated to low frequency alleles, focusing on SNPs with common alleles reduces the proportion of wrong SNPs in the data, and consequently their influence on summary statistics. In our application to cattle NGS data, this strategy was efficient because wrong SNP calls were the only detectable effect of sequencing errors on the data. In particular, genotyping errors at true SNP calls had no impact on the summary statistics, as shown by the perfect match between summary statistics computed from NGS data or genotyping data at the 800K chip positions (S20 Fig). Indeed, NGS genotypes had been corrected by imputation, taking advantage of the large sample size and / or sequencing depth within each breed [40]. As this might not be the case in all data sets, other strategies could be applied to correct for sequencing errors, while keeping the main idea of an ABC approach based on AFS and LD statistics. For instance, one could simulate NGS data with the same coverage and error rates as the

observed data, rather than perfect genotype data, and compute observed and summary statistics directly from raw NGS data, using dedicated algorithms that account for the uncertainty of genotype calls. Such algorithms are available both for AFS [55] and LD statistics [56], which is another advantage of using these standard summary statistics. However, this strategy would be much more computationally demanding than the one we used here.

Contribution to the demographic history of cattle

Until recently, effective population size estimations in cattle, and more generally in all livestock species, were mostly based on two approaches. The first approach focuses on the few most recent generations and estimates population size from the increase of inbreeding or coancestry along generations, based on pedigree or molecular information [17–19]. Using this approach, population size estimations from around 50 animals in Holstein to around 150 animals in Simental (closely related to Fleckvieh) were obtained [19]. These estimated populations sizes are qualitatively consistent with ours, as we estimated that the recent population size in Fleckvieh was about three fold larger than in Holstein, but the actual values obtained with these approaches were substantially lower than our estimates (790 in Holstein and 2,220 in Fleckvieh). This may partly be due to the small bias observed with our approach in the simulated decline scenario, using either the median (Fig 3) or the mode (S10 Fig) of the posterior distribution as point estimation. But it is also important to mention that the animals sequenced in the 1,000 bull genomes project were chosen among key ancestors of the breed, so the most recent population size estimated in our study might reflect the population size a few generations ago rather than the current one. This could partly explain the discrepancy between the estimates, because artificial selection has been particularly intensive within the few last generations, leading to a further decline of effective population size.

The second approach is based on the average r^2 over different bins of genetical distance [26, 47], which has been already mentioned earlier in the discussion. It aims at estimating population size on a much larger time scale and has been extensively applied in cattle [41, 57, 58] and other livestock species [59]. Indeed, a very large number of animals have been genotyped using SNP chips in these species, sometimes for other purposes, such as QTL detection, and used for LD estimation. In addition to the methodological issues related to this approach, the use of SNP chip data for population size estimation presents its own limitations. The ascertainment bias associated to SNP chip data does not only influence AFS statistics but also LD statistics, which in turn affects population size estimation. This is outlined by the fact that our ABC approach based only on LD summary statistics infers different population size histories when these statistics are computed from all the SNPs found by NGS, or only from those that overlap with the 800K chip (S30 Fig). Regrettably, this influence of ascertainment bias on population size estimations obtained from LD statistics is generally not accounted for by the studies using LD. Besides, considering LD alone leads to a different prediction than considering LD and AFS together (S30 Fig), and our simulation results suggest that the former prediction is less reliable. Overall, the use of NGS data, and of dedicated inference approaches taking advantage of these data, should thus considerably improve our understanding of livestock evolutionary history, at least above 600 generations (3,000 years in cattle) BP (S30 Fig).

To our knowledge, the first (and so far the only) estimation of population size history in cattle based on NGS data was obtained by [12]. This result was based on the distribution of IBS segment length in one Australian Holstein bull sequenced at 13X coverage. The overall histories found in this study and in ours are quite consistent, as they both exhibit a strong decline of population size from about 20,000 years BP to the very recent past, but our estimations of population size are generally larger. For instance the population size before this decrease was

around 20,000 in their study and around 50,000 in ours, and the population size 1,000 years ago was around 2,000 in their study and around 4,000 in ours. Although the most obvious difference between the two approaches is that they use different summary statistics, ABC estimations obtained from IBS statistics rather lead to larger or equal population sizes than those obtained from AFS and LD statistics (S31 Fig). Thus, we think that the difference between our estimation and that in [12] more likely comes from a difference in the recombination rate. This rate is set to 1e-8 per generation and per bp in [12], while our approach would rather provide an estimation around 4e-9. Assuming that our estimation is correct, the overestimation of r by a factor two in [12] could lead to an underestimation of N by the same factor, because one essential parameter determining the IBS segment length distribution is the scaled recombination rate $2Nr$. Further work will be needed to better understand the difference between the two estimations.

Perspectives

Our ABC approach, as well as other SMC [9, 10] or IBS based methods [12], assumes that the considered population has evolved forever as an isolated population. This is obviously a strong hypothesis: for instance the cattle breeds considered here have actually diverged from a common ancestral population. Several studies have demonstrated that population structure can leave genomic signatures similar to those of population size changes, even if each of the subpopulations is actually of constant size [60–64]. Consequently, population size histories estimated by single population approaches should be interpreted with caution. However, we anticipate that our study will pave the way for future approaches inferring population size histories jointly in multiple populations, while accounting for the history of divergences and migrations in these populations. ABC represents a perfect framework for developing such approaches, because of the flexibility offered by the simulation procedure. It is already widely used in population genetics for estimating parameters in multiple population models including for instance admixture events and some population size changes [65]. Besides, previous studies showed that structured models and population size change models can be distinguished using ABC [61].

In this study, the flexibility offered by ABC allowed us to infer parameters under the true coalescent with mutation and recombination, rather than under the SMC approximation as in [9, 10, 54]. One could actually go much further and relax also the hypotheses of the Kingman's (1982) coalescent. For instance in cattle, genealogies in the most recent generations are highly unbalanced, because a few bulls with outstanding genetic values have been used to produce thousands of offsprings through artificial insemination. Such genealogies are not consistent with the Kingman's coalescent, but specific algorithms combining the Kingman's coalescent with a few generations of forward-in-time simulations could certainly be implemented and used to perform ABC estimations in this context.

Methods

The ABC approach

Assume we observe a dataset $\mathcal{D}$, from which we want to estimate the parameters θ of a given model. In a Bayesian framework, this involves computing the posterior probability $\mathbb{P}(\theta \mid \mathcal{D})$ for any possible parameter value. In many situations, and in particular in population genetics, this posterior cannot be derived because of the model complexity and even numerical evaluations are impossible due to the high dimensionality of the observed data space. The idea of ABC [20] is to replace in this context the full dataset $\mathcal{D}$ by a vector of summary statistics $\mathcal{S}$ capturing most information contained in the data and to estimate model parameters based on the

approximate posterior $\mathbb{P}(\theta \mid \mathcal{S})$. The estimating procedure consists in sampling a very large number of parameter values from a prior distribution, simulating datasets from these parameter values, and accepting the parameter values leading to summary statistics that are sufficiently similar to those of the observed dataset.

Several strategies can then be used to estimate the posterior distribution. The easiest one, called rejection, is to compute the empirical distribution of the accepted parameter values. To account for the imperfect match between accepted and observed summary statistics, accepted parameter values can also be adjusted by various regression methods, using the associated summary statistics as explanatory variables. The general idea of these methods is to assume a local regression model in the vicinity of S, with an equation of the form

$$\theta_k = m(S_k) + \epsilon_k \quad (2)$$

where θ_k is the value of parameter θ in the kth simulated sample, S_k is the vector of summary statistics in this sample, $m()$ is a regression function varying between approaches, and ϵ_k is a random noise. This model is fitted using all accepted samples. Adjusted parameter values are then obtained by

$$\hat{\theta}_k = \hat{m}(S) + \hat{\epsilon}_k$$

where $\hat{m}$ is the estimated regression function and $\hat{\epsilon}_k$ is the empirical noise, and the posterior distribution is finally computed as the empirical distribution of these adjusted values. A general review on these aspects can be found in [46].

Model and priors

Here the observed data $\mathcal{D}$ is a set of n diploid genomes sampled from a single panmictic population, and the model assumed to have generated these data is the coalescent with mutation and recombination [16]. We assume that effective population size varied according to a piecewise constant process. Following [9, 10], we considered a fixed number of time windows, whose size increased exponentially from recent to old periods. More precisely, we used $I = 21$ windows of the form $[t_i, t_{i+1}]$, where $t_i = \exp(\log(1 + aT)i/(I-1)) - 1)/a$ generations BP for i from 0 to $I-1$, with $T = 130{,}000$ and $a = 0.06$, and $t_I = +\infty$. These specific values of T and a were chosen to capture important periods of cattle history. Modifying T would allow population size changes to occur on a longer or shorter period in the past, and modifying a would allow to describe more precisely one specific part of the history, playing on the ratio between the length of recent versus old time windows. With our parametrization, the most recent time window ranged from present to 10 generations BP, the second most recent ranged from 10 to 25 generations BP, . . .the second oldest ranged from 83,000 to 130,000 generations BP and the oldest included all generations above 130,000 generations BP.

The parameters of this model are the population sizes N_i for i from 0 to $I-1$, the per generation per site recombination rate r and the per generation per site mutation rate μ. Prior distributions for the population sizes were taken uniform in the log 10 scale, from 10 to 100,000. In order to avoid unrealistic trajectories, we also set that the ratio of population sizes between two consecutive time windows could not exceed 10. In practice, we thus sampled $\log_{10}(N_0)$ uniformly between 1 and 5, and iteratively computed $\log_{10}(N_i) = \max(\min(\log_{10}(N_{i-1}) + \alpha, 5), 1)$, with α sampled uniformly between −1 and 1. For the recombination rate, we used an uniform prior between 1e-9 and 1e-8, consistent with recent estimations in cattle [66]. For the mutation rate we considered a fixed value, in order to compare our estimation approach with other recent ones making the same hypothesis [9, 10, 12], but it would be straightforward to use a prior distribution instead. This value was taken equal to 1e-8, as in [12].

Summary statistics

We summarized each sample of n diploid genomes using a combination of statistics related to the allele frequency spectrum (AFS) and the average linkage disequilibrium (LD) over the genome. AFS statistics included the overall proportion of polymorphic sites over the genome (one statistic) and, among these polymorphic sites, the proportion of those with i copies of the minor allele, for i from 1 to n (n statistics). LD statistics included the average r^2 over 18 different sets of SNP pairs (18 statistics), where each set was characterized by a different physical distance between SNPs. Indeed, the expected value of r^2 between two SNPs at genetic distance c is related to the population size $1/2c$ generations BP [26]. Thus, for each of the time windows of our model, we computed r^2 for SNP pairs whose physical distance would approximately correspond to a genetic distance of $1/2t$ (± 5%), where t was the middle of the window, assuming a recombination rate of 1.0 cM/Mb. For the two most recent windows, the physical distance between SNPs derived from this formula was larger than 2Mb, which could not be achieved in our simulations (see below). We thus considered only 19 statistics out of 21 windows, corresponding to distances between SNPs going from 282 bp to 1.4 Mb. We further dropped the LD statistic corresponding to a distance of 282 bp, both in the simulation study and in the real data analysis, because with our cattle data (described below) it had a strikingly low value, which was likely due to a technical problem related to the sequencing, the calling or the accuracy of the assembly. Consequently, the smallest distance bin used in our study was finally equal to 470 bp. This is specific to our study and smaller distances might be used in future studies. By default, the r^2 computed between two SNPs was the zygotic LD, i.e. the correlation between the vectors of n genotypes observed at the two SNPs [67]. But for some comparative analyzes we also calculated the well-known gametic LD, where the correlation is computed between the two vectors of $2n$ alleles observed at the two SNPs. Note that this second option is only possible for haploid or phased data.

In many situations, we computed these summary statistics only from SNPs above a given minor allele frequency (MAF) threshold, whose value could differ between AFS and LD statistics. For a MAF threshold corresponding to c copies of the minor allele, the overall proportion of SNPs was changed to the overall proportion of SNPs with more than c copies of the minor allele, and all other proportions in the AFS were computed relative to SNPs with more than c copies of the minor allele. Overall, only $n + 2 - c$ statistics were available in this case, instead of $n + 1$ without MAF threshold. In contrast, the number of LD statistics was not affected by the MAF threshold.

In a few specific analyses, we also computed summary statistics related to the distribution of IBS segment length. We summarized this distribution by a set of 11 quantiles, from 0.0001 to 1 − 0.0001.

Implementation

We simulated 250,000 samples of 100 haploid genomes using *ms* [68], with parameters sampled from the priors described above. We chose this software because it allows simulating the exact coalescent with mutation and recombination, but faster algorithms based on approximations of this model could be used in future studies. For computational reasons, each haploid genome included only 100 independent 2Mb-long long segments. From each simulated sample of 100 haploid genomes, five different samples of n diploid genomes were created, for n equal to 10, 15, 20, 25 and 50. Each of these samples was created by choosing at random $2n$ haploid genomes among 100 (without replacement). In addition, 200,000 samples of 25 diploid genomes were simulated directly from *ms* samples of 50 haploid genomes. Thus, ABC analyses focusing on a sample size of 25 diploid genomes were based on 450,000 simulated samples (unless specified), while analyses involving other sample sizes were based on 250,000 simulated samples.

For the real data set, a total of 234 phased bull genomes were obtained from the 1,000 bull genomes project, Run II [40]. These included 129 Holstein (125 Black and 4 Red), 43 Fleckvieh, 47 Angus and 15 Jersey animals. Holstein animals came from various flocks with distinct geographical origins. In order to study homogeneous groups, we thus focused on the 52 Holstein animals from Australia (other geographical origins had significantly lower sample sizes). We further selected 25 unrelated animals within each breed with the following procedure: first, we removed all animals that were either extremely inbred or extremely related to another sampled animal, based on the genomic relationship matrix computed from GCTA [69]. Then, we sampled 25 animals at random among the remaining ones. For the Jersey breed, as only 15 animals were available and as they were found to be all unrelated to each other, we kept them all.

The summary statistics described above were computed using the same python script for both simulated and cattle samples. Since the length of cattle chromosomes was much larger than that of simulated segments (2Mb), we first cut each cattle chromosome into consecutive but non-overlapping 2Mb-long segments. To keep the approach computationally efficient, the average LD for a given distance bin was not evaluated from all SNP pairs satisfying the distance condition, but from a random subset of these pairs. This subset was selected by an iterative search along each 2Mb-long region, so that intervals defined by all SNP pairs did not overlap.

With the default parameter values described above, simulating 100 genomic samples and computing all summary statistics for these samples took approximately three hours on a standard computer, using a single core. Using 200 cores in parallel on a computing cluster, we could obtain 450,000 samples of summary statistics in less than 48 hours.

The final ABC estimation, based on the comparison of the simulated and observed summary statistics, was performed in R using the package *abc* [70]. By default, we accepted simulated samples with a tolerance rate of 0.005 and adjusted accepted values by a neural network regression approach [32]. This approach allows to reduce the dimension of the set of summary statistics and accounts for the non-linearity of the regression function m linking parameters and statistics (Eq (2)). Neural network regression was applied with the default parameter values of the function *abc*, except for the final analysis of all cattle breeds where 100 (instead of 10) neural networks were fitted in order to get more stable estimations. For each parameter, a point estimate was obtained by taking the median of the posterior distribution. Variations from this default strategy were also tried, as mentioned in the results section. In particular, we also estimated posterior distributions using rejection or ridge regression [33], using the default values implemented in the *abc* package.

Cross validation analyzes

We evaluated the performance of ABC using several subsets of summary statistics and several choices of MAF threshold, sample size, estimation approach, or tolerance. For each specific combination of these parameters, we conducted a cross validation study based on $K = 2000$ simulated samples, using the R function *cvabc*. The prediction error (PE) associated to a given parameter value θ was computed as $(1/K)(\sum_{k=1}^{2000}(\hat{\theta}_k - \theta_k^*)^2)/var(\theta)$, where θ_k^* is the true value of θ in the kth simulated sample, $\hat{\theta}_k$ is the point estimation of this value provided by ABC, and $var(\theta)$ is the prior variance of θ. With this scaling, estimating θ_k from the prior distribution of θ would result in a PE of 1.

Similarly, the estimation bias for θ was computed as $(1/K)\sum_{k=1}^{2000}(\hat{\theta}_k - \theta_k^*)$, and the empirical coverage of the 90% credible interval was evaluated by $(1/K)\sum_{k=1}^{2000} 1(q_{10}(\theta_k) \leq \theta_k^* \leq q_{90}(\theta_k))$, where $q_{10}(\theta_k)$ and $q_{90}(\theta_k)$ are the 5% and 95% quantiles of the posterior distribution of θ_k, and $1(C)$ is the indicative function equal to 1 if condition C is satisfied and 0 otherwise.

When computing these metrics for the population size N in a given time window, we focused on parameter $\theta = \log_{10}(N)$ rather than $\theta = N$. Without this rescaling, PEs and biases would only reflect the estimation accuracy for large populations, while estimation errors concerning small populations would be masked.

Rescaling time from generations to coalescent units

Considering a population with variable population size, let $N(t)$ be the haploid population size at generation t and $\tau = \frac{t}{N(0)}$ be a rescaling of time in $N(0)$ units. In this time scale, the history of population size changes is summarized by the function:

$$f(\tau) = \frac{N(\tau)}{N(0)}, \quad \tau \geq 0$$

It can be shown [71] that the genealogical process of a sample of size n from this population, and in particular the joint distribution of all coalescence times, is identical to the genealogical process of a sample of size n in a constant size population where time would be rescaled by the function

$$\Lambda(\tau) = \int_0^\tau \frac{1}{f(x)} dx$$

Consequently, all variable population size histories can be related to the classical Kingman's coalescent. In this process, the expected TMRCE in a sample of size n is $\frac{2}{n(n-1)}$ and the expected TMRCA is $2(1 - 1/n)$.

In S2 Fig, the PE obtained in time window $[t_i, t_{i+1}]$ for a given population size history was allocated to the rescaled interval $[u_i, u_{i+1}]$. Applying the equations above to the specific situation of a piecewise constant population size process, u_i was computed as

$$u_i = \sum_{k=0}^{i} \frac{\tau_k - \tau_{k-1}}{f_k}, \quad i \geq 1$$

with $\tau_k = \frac{t_k}{2N_0}$ and $f_k = \frac{N_k}{N_0}$. PE were then averaged over histories, for several values of u between 1e-5 and 100. Note that N_0 is the haploid population size here, while the population sizes mentioned anywhere else in this paper are always diploid population sizes. We used $\tau_k = \frac{t_k}{2N_0}$, instead of the classical $\tau_k = \frac{t_k}{N_0}$ mentioned above, in order to get an expected TMRCA approximately equal to 1 (rather than 2) for large samples, which facilitates the reading of S2 Fig.

Additional simulated datasets

Twenty scenarios with fixed population size history were considered for validation, see Fig 3, S11, S13, and S14 Figs. For each of these scenarios, 20 PODs were simulated. Each of them included 25 diploid genomes and 500 independent 2Mb-long segments. Population size parameters were the same in all 20 replicates of each scenario, and the per site recombination rate was also constant and equal to 5e-9.

Comparison of summary statistics obtained from NGS and genotyping data

For 12 of the 129 Holstein bulls considered in this study, genotypes on the 800K Illumina bovine SNP chip were obtained from the Gembal project [72]. Among the 708,771 SNPs retained in this study after quality control, 562,746 were polymorphic among the 12 bulls

considered here. These SNPs were used to compute the polymorphic site AFS and the LD summary statistics from genotyping data. The rate of false negative SNPs in the NGS data was estimated by the proportion of these 562,746 positions for which no SNP was called from the NGS data. Similarly, the rate of false positive SNPs in these NGS data was estimated by considering the 145,978 SNP positions that were found monomorphic with the 800K genotypes, and computing the proportion of these positions where a SNP was called in the NGS data.

Software and data availability

Python and R scripts for the PopSizeABC method can be found at https://forge-dga.jouy.inra.fr/projects/popsizeabc/. Simulated and observed summary statistics used in this study are also provided on this web page.

Supporting Information

S1 Fig. Accuracy of credible intervals obtained by ABC. Empirical coverage (left) and width (right) of the 90% credible interval for the population size in each time window. The empirical coverage is the proportion of simulated histories for which the true population size was included in the 90% credible interval of the posterior distribution. If the posterior distribution was correctly estimated, this proportion should have been 90%, as shown by the black horizontal solid line. Parameter settings were the same as in Fig 1.
(PDF)

S2 Fig. Accuracy of ABC estimation along the coalescent process. Prediction error for the estimated population size when time is measured in units of the expected time to the most recent common ancestor (TMRCA) of the sample. Prediction errors were evaluated from 2,000 random population size histories. Black vertical dotted lines indicate the expected time to the most recent coalescence event, $E[TMRCE]$, and the expected TMRCA, $E[TMRCA]$. Summary statistics considered in the ABC analysis were (i) the AFS and (ii) the average zygotic LD for several distance bins. These statistics were computed from $n = 25$ diploid individuals, using all SNPs for AFS statistics and SNPs with a MAF above 20% for LD statistics. The posterior distribution of each parameter was obtained by neural network regression [32], with a tolerance rate of 0.005. Population size point estimates correspond to the median of the posterior distribution.
(PDF)

S3 Fig. Accuracy of credible intervals obtained by ABC and relative importance of the summary statistics. Empirical coverage (left) and width (right) of the 90% credible interval for the population size in each time window. Parameter settings were the same as in Fig 2. The very large credible intervals obtained on average with AFS statistics, in some time windows, are due to a retalively small number of PODs with extreme values.
(PDF)

S4 Fig. Accuracy of ABC estimation based on LD summary statistics. Prediction error for the estimated population size in each time window, evaluated from 2,000 random population size histories. Summary statistics considered in the ABC analysis were the average gametic LD (triangles) or the average zygotic LD (circles) for several distance bins. These statistics were computed from $n = 25$ diploid individuals, using different MAF thresholds. Other parameter settings were the same as in Fig 2.
(PDF)

S5 Fig. Influence of the number of simulated data sets on ABC estimation. Top: Prediction error for the estimated population size in each time window (left) and standard deviation of this error (right). Bottom: Empirical coverage (left) and width (right) of the 90% credible interval for the population size in each time window. These quantiles were evaluated from 2,000 random population size histories. For each of these histories, one POD of $n = 25$ diploid genomes was simulated, where each genome consisted in 100 independent 2Mb-long segments. Population size history was estimated from this POD by ABC, for various numbers of simulated datasets (see the legend) with the same sample size ($n = 25$) and genome length (100 independent 2MB segments). Summary statistics considered in the ABC analysis were (i) the AFS and (ii) the average zygotic LD for several distance bins. AFS statistics were computed using all SNPs and LD statistics were computed using SNPs with a MAF above 20%. The posterior distribution of each parameter was obtained by neural network regression, with the tolerance rate leading to the smallest prediction error. Population size point estimates were obtained from the median of the posterior distribution.
(PDF)

S6 Fig. Influence of the genome length of simulated and observed data sets on ABC estimation. Top: Prediction error for the estimated population size in each time window (left) and standard deviation of this error (right). Bottom: Empirical coverage (left) and width (right) of the 90% credible interval for the population size in each time window. These quantiles were evaluated from 2,000 random population size histories. For each of these histories, one POD of $n = 25$ diploid genomes was simulated, where each genome consisted in 10, 50 or 100 independent 2Mb-long segments (see the legend). Population size history was estimated from this POD by ABC, using 450,000 simulated datasets with the same sample size ($n = 25$) and genome length. The posterior distribution of each parameter was obtained by neural network regression, with a tolerance rate of 0.005. All other settings are similar to S5 Fig.
(PDF)

S7 Fig. Using different genome lengths for simulated and observed data sets. Top: Prediction error for the estimated population size in each time window (left) and standard deviation of this error (right). Bottom: Empirical coverage (left) and width (right) of the 90% credible interval for the population size in each time window. These quantiles were evaluated from 2,000 random population size histories. For each of these histories, one POD of $n = 25$ diploid genomes was simulated, where each genome consisted in 10 or 100 independent 2Mb-long segments (see the legend). Population size history was estimated from this POD by ABC, using 450,000 simulated datasets with the same sample size ($n = 25$) but a possibly different genome length (see the legend). The posterior distribution of each parameter was obtained by neural network regression, with a tolerance rate of 0.005. All other settings are similar to S5 Fig.
(PDF)

S8 Fig. Influence of the sample size on ABC estimation. Top: Prediction error for the estimated population size in each time window (left) and standard deviation of this error (right). Bottom: Empirical coverage (left) and width (right) of the 90% credible interval for the population size in each time window. These quantiles were evaluated from 2,000 random population size histories. For each of these histories, one POD of n diploid genomes was simulated, for different values of n between 10 and 50 (see the legend). Each genome consisted in 100 independent 2Mb-long segments. Population size history was estimated from this POD by ABC, using 450,000 simulated datasets with the same sample size and genome length. All other settings are similar to S5 Fig.
(PDF)

S9 Fig. Influence of MAF threshold on ABC estimation. Top: Prediction error for the estimated population size in each time window (left) and standard deviation of this error (right). Middle: Bias for the estimated population size in each time window. Bottom: Empirical coverage (left) and width (right) of the 90% credible interval for the population size in each time window. These quantiles were evaluated from 2,000 random population size histories. For each of these histories, one POD of $n = 25$ diploid genomes was simulated, where each genome consisted in 100 independent 2Mb-long segments. Population size history was estimated from this POD by ABC, using 450,000 simulated datasets with the same sample size and genome length. Summary statistics considered in the ABC analysis were (i) the AFS and (ii) the average zygotic LD for several distance bins. AFS statistics were computed using different MAF thresholds, LD statistics were computed from SNPs with a MAF above 20%. The posterior distribution of each parameter was obtained by neural network regression, with a tolerance rate of 0.005. Population size point estimates were obtained from the median of the posterior distribution.
(PDF)

S10 Fig. Estimation of population size history from the mode of the posterior distribution in six different simulated scenarios. All settings are similar to Fig 3, except that population size point estimates were obtained from the mode of the posterior distribution.
(PDF)

S11 Fig. Estimation of population size history in the zigzag scenario and five related scenarios. a scenario where all population sizes are divided by ten compared to the original zigzag ("zigzag small", top right), a scenario where only the recent bottleneck of the original zigzag is simulated ("bottleneck1 recent large", middle left), a scenario corresponding to the history wrongly inferred by ABC based on data from the "bottleneck1 recent large" scenario (middle right), and two scenarios where only the recent (bottom left) or the old (bottom right) bottleneck of the "zigzag small" are simulated. All settings are similar to Fig 3.
(PDF)

S12 Fig. Observed and best simulated summary statistics in the "bottleneck1 recent large" scenario. For one of the five PODs analyzed in this scenario, observed AFS (left) and LD (right) statistics are shown by green full circles. The average value of these statistics over the five best simulated data sets, i.e. the five simulated data sets leading to the smallest distance between observed and simulated statistics, are shown by blue crosses. The variation of these statistics over the five best simulated data sets is also indicated by blue dotted lines, which correspond to the average value plus (or minus) twice the standard deviation of each statistic.
(PDF)

S13 Fig. Estimation of population size history in four scenarios including a bottlenck followed by a population decline. Population size varied between 60,000 and 6,000 individuals in the top panels, and between 6,000 and 600 individuals in the bottom panels. Population size changes occurred between 2,300 and 50 generations BP in the left panels, and between 34,000 and 900 generations BP in the right panels. All settings are similar to Fig 3.
(PDF)

S14 Fig. Estimation of population size history in the decline scenario and five related scenarios. a sudden (rather than continuous) decline from 40,000 to 300 individuals occurring 200 generations BP (top right), a sudden decline from 40,000 to 300 individuals occurring 1,000 generations BP (middle left), the same sudden decline followed by an expansion to 5,000 individuals occurring 580 generations BP (middle right) or an expansion to 1,000 individuals occurring 140 generations BP (bottom left), and a scenario similar to the continuous decline (top left) but

including a sudden decline to 100 individuals between 230 and 140 generations BP, followed by an expansion to 1,000 individuals (bottom right). All settings are similar to Fig 3.
(PDF)

S15 Fig. Estimation of past effective population size using MSMC with four haplotypes in six different simulated scenarios. For each scenario, the five PODs considered for MSMC estimation were the same as in Fig 3. The expected TMRCA shown here is also the same as in Fig 3, it corresponds to samples of 50 haploid sequences.
(PDF)

S16 Fig. Estimation of past effective population size using MSMC with eight haplotypes in six different simulated scenarios. For each scenario, the five PODs considered for MSMC estimation were the same as in Fig 3. The expected TMRCA shown here is also the same as in Fig 3, it corresponds to samples of 50 haploid sequences.
(PDF)

S17 Fig. Estimation of past effective population size using MSMC with four haplotypes in the decline scenario and five related scenarios. For each scenario, the five PODs considered for MSMC estimation were the same as in S14 Fig. The expected TMRCA shown here is also the same as in S14 Fig, it corresponds to samples of 50 haploid sequences.
(PDF)

S18 Fig. Estimation of past effective population size using MSMC with eight haplotypes in the decline scenario and five related scenarios. For each scenario, the five PODs considered for MSMC estimation were the same as in S14 Fig. The expected TMRCA shown here is also the same as in S14 Fig, it corresponds to samples of 50 haploid sequences.
(PDF)

S19 Fig. Influence of phasing errors on MSMC estimation. Estimation of past effective population size using MSMC with four haplotypes in the "small" scenario (top), the "decline" scenario (middle) and the "expansion" scenario (bottom). MSMC analyzes were run from perfectly phased data, phased data with 1 or 10 switch errors per Mb and diploid individual, or unphased data (i.e. two unphased diploid individuals). All other settings are similar to S15 Fig.
(PDF)

S20 Fig. Comparison of summary statistics obtained from NGS and genotyping data. polymorphic site AFS, i.e. without the overall proportion of SNPs (left), average gametic LD (middle) and average zygotic LD (right). These statistics were computed from 12 Holstein animals for which both NGS data and genotyping data were available, using only SNP positions from the 800K chip (even for the NGS data statistics). No MAF threshold was used.
(PDF)

S21 Fig. False positive and false negative rates of SNP detection in the 1,000 bull genomes project. Error rates were computed from 12 Holstein animals for which both NGS data and genotyping data were available. False positive SNPs were positions that were found polymorphic in the NGS data but not in the 800K data. Their minor allele count in the NGS data was called the wrong minor allele count. False negative SNPs were positions that were found polymorphic in the 800K data but not in the NGS data. Their minor allele count in the 800K data was called the true minor allele count.
(PDF)

S22 Fig. Estimation of population size history using ABC without rare SNPs in five different simulated scenarios. All settings are similar to Fig 3, except that AFS statistics were

computed only from SNPs with a MAF above 20%.
(PDF)

S23 Fig. Ninety percent credible intervals of estimated population size history in four cattle breeds. Holstein (top left), Angus (top right), Fleckvieh (bottom left) and Jersey (bottom right). Parameter settings are the same as in Fig 6.
(PDF)

S24 Fig. Predictive posterior check of the population size history estimated in the Holstein cattle breed (Fig 6). Ten thousand genomic samples were simulated under population size histories that were sampled from the posterior distribution estimated in Fig 6. Four combinations of summary statistics were computed from each sample: AFS and LD statistics (top left), AFS statistics alone (top right), LD statistics alone (bottom left) and IBS statistics (bottom right, see the Methods for a detailed description of these statistics). For each of these combinations, a principal component analysis (PCA) of the 10,000 simulated samples was performed: the projection of all samples on the two first dimensions of this PCA are plotted in black. The vector of summary statistics observed in Holstein was then projected on the same hyperplan. It always fell within the cloud of simulated summary statistics, which shows that the estimated history is able to reproduce summary statistics that are indeed similar to the observed ones. Interestingly, this also holds for IBS statistics, which were not used for the estimation. Results are shown for the Holstein breed but they were similar for the other breeds.
(PDF)

S25 Fig. Observed and best simulated summary statistics in the Holstein cattle breed. Observed AFS (left) and LD (right) statistics are shown by green full circles. The average value of these statistics over the five best simulated data sets, i.e. the five simulated data sets leading to the smallest distance between observed and simulated statistics, are shown by blue crosses. The variation of these statistics over the five best simulated data sets is also indicated by blue dotted lines, which correspond to the average value plus (or minus) twice the standard deviation of each statistic.
(PDF)

S26 Fig. Influence of population size changes on LD statistics. LD statistics for several scenarios inplying a sudden expansion from 500 to 50,000 individuals (left) or a sudden decline from 50,000 to 500 individuals (right). Several expansion or decline times were considered, as well as two scenarios with a constant population size of 500 or 50,000 individuals (see the legend). For each scenario, LD statistics were averaged over 20 PODs including 25 diploid genomes and 100 2Mb-long regions. In contrast with expansion scenarios, some decline scenarios lead to even larger LD statistics than those obtained for a constant small population. Indeed, as these declines are very old compared to the expected TMRCA of a population of 500 individuals, their main effect is to increase, at some loci, the time during which the sample has only two ancestral lineages. Because this increase is very large (backward in time, population size, and thus expected coalescence time, are suddenly multiplied by 100), mutations occuring in this part of the coalescence tree eventually represent a large proportion of all oberved polymorphic sites. Besides, for two linked loci with similar topologies of the coalescence tree, mutations occuring in this part of the tree lead to very high r^2 values, up to 1 if the topologies are exactly the same.
(PDF)

S27 Fig. Accuracy of ABC and relative importance of LD and AFS in different families of scenarios. Prediction error for the estimated population size in each time window, focusing on

scenarios with a population size below 1,000 (top left), above 10,000 (top right), below 1,000 in the last 200 generations and above 10,000 for times more ancient than 13,000 generations BP (bottom left) or above 10,000 in the last 200 generations and below 1,000 for times more ancient than 13,000 generations BP (bottom left). For the two latter scenarios, the time window where population size goes from above 10,000 to below 1,000 (or vice versa) is delimited by vertical dotted lines. For each scenario category, PE were evaluated from 2,000 random histories. Summary statistics considered in the ABC analysis were either the AFS statistics alone, the LD statistics alone or the AFS and LD statistics together (see the legend). All other settings are similar to Fig 2.
(PDF)

S28 Fig. Estimation of population size history using different ABC settings in the "bottleneck1 old large" scenario. Summary statistics considered in the ABC analysis were either the AFS statistics alone (left column), the LD statistics alone (middle column), or the AFS and LD statistics together (right column). AFS statistics were computed using either all SNPs (top panels) or only those with a MAF above 20% (bottom panels). All other settings are similar to Fig 3.
(PDF)

S29 Fig. Accuracy of ABC estimation based on the distribution of IBS segment lengths. Prediction error for the population size in each time window, evaluated from 2,000 random population size histories. Summary statistics considered in the ABC analysis included several combinations of (i) the AFS, (ii) the average zygotic LD for several distance bins and (iii) the distribution of IBS segment lengths within one diploid individual. These statistics were computed from $n = 25$ diploid individuals, using all SNPs for AFS and IBS statistics and SNPs with a MAF above 20% for LD statistics. Other parameter settings are the same as in Fig 2.
(PDF)

S30 Fig. Added value of NGS for population size history estimation. Estimation of population size history in the Holstein cattle breed using ABC, based on whole genome NGS data from $n = 25$ animals. Summary statistics considered in the ABC analysis included different combinations of (i) the AFS and (ii) the average zygotic LD for several distance bins. These statistics were computed either from the SNPs that are included in the 800K SNP chip or from all SNPs found in the NGS data. A MAF threshold of 20% was used for all curves and statistics. Other parameter settings are the same as in Fig 5.
(PDF)

S31 Fig. Population size history in Holstein using IBS statistics. Estimation of population size history in the Holstein cattle breed using ABC, based on whole genome NGS data from $n = 25$ animals. Summary statistics considered in the ABC analysis were either both the AFS and the average zygotic LD for several distance bins, or the distribution of IBS segment lengths within one diploid individual. These statistics were computed using SNPs with a MAF above 20%. Other parameter settings are the same as in Fig 5.
(PDF)

Acknowledgments

Data analyzes were performed on the computer cluster of the Genotoul bioinformatics platform Toulouse Midi-Pyrénées (www.bioinfo.genotoul.fr). The cattle genomes were obtained from the 1,000 bull genomes project (www.1000bullgenomes.com). The 12 Holstein bull genotypes for the bovine 800K SNP chip were obtained from the GEMBAL project, which is funded by the Agence Nationale de la Recherche (ANR-10-GENM-0014), APISGENE, Races de

France and INRA "AIP Bioressources". Preliminar analyzes were performed by Stanislas Sochacki, who was funded by the Projets Exploratoires Pluridisciplinaires (PEPS 2012 Bio-Maths-Info). Flora Jay's position was funded by the Agence Nationale de la Recherche Demochips project (ANR-12-BSV7-0012). We would like to thank Lounès Chikhi, Olivier Mazet, Simona Grusea, Bertrand Servin, Didier Boichard and all members of the Demochips project for their useful comments on this study. We are also grateful to Bertrand Servin for sharing some Python code, and to three anonymous reviewers for their constructive comments on previous versions of the manuscript.

Author Contributions

Conceived and designed the experiments: SB WR FJ SM FA. Analyzed the data: SB WR. Contributed reagents/materials/analysis tools: SB WR FJ. Wrote the paper: SB WR FJ SM FA.

References

1. Lorenzen E, Nogues-Bravo D, Orlando L, Weinstock J, Binladen J, Marske K, et al. Species-specific responses of Late Quaternary megafauna to climate and humans. Nature. 2011; 479(7373):359–364. doi: 10.1038/nature10574 PMID: 22048313
2. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. Interrogating a High-Density SNP Map for Signatures of Natural Selection. Genome Research. 2002; 12(12):1805–1814. Available from: http://genome.cshlp.org/content/12/12/1805.abstract. doi: 10.1101/gr.631202 PMID: 12466284
3. Goldstein DB, Chikhi L. HUMAN MIGRATIONS AND POPULATION STRUCTURE: What We Know and Why it Matters. Annual Review of Genomics and Human Genetics. 2002; 3(1):129–152. Available from: http://dx.doi.org/10.1146/annurev.genom.3.022502.103200. doi: 10.1146/annurev.genom.3.022502.103200 PMID: 12142358
4. Quéméré E, Amelot X, Pierson J, Crouau-Roy B, Chikhi L. Genetic data suggest a natural prehuman origin of open habitats in northern Madagascar and question the deforestation narrative in this region. Proceedings of the National Academy of Sciences. 2012; 109(32):13028–13033. Available from: http://www.pnas.org/content/109/32/13028.abstract. doi: 10.1073/pnas.1200153109
5. Pybus OG, Rambaut A, Harvey PH. An Integrated Framework for the Inference of Viral Population History From Reconstructed Genealogies. Genetics. 2000; 155(3):1429–1437. Available from: http://www.genetics.org/content/155/3/1429.abstract. PMID: 10880500
6. Ho SYW, Shapiro B. Skyline-plot methods for estimating demographic history from nucleotide sequences. Molecular Ecology Resources. 2011; 11(3):423–434. Available from: http://dx.doi.org/10.1111/j.1755-0998.2011.02988.x. doi: 10.1111/j.1755-0998.2011.02988.x PMID: 21481200
7. Burgarella C, Navascuès M, Zabal-Aguirre M, Berganzo E, Riba M, Mayol M, et al. Recent population decline and selection shape diversity of taxol-related genes. Molecular Ecology. 2012; 21(12):3006–3021. Available from: http://dx.doi.org/10.1111/j.1365-294X.2012.05532.x. doi: 10.1111/j.1365-294X.2012.05532.x PMID: 22574693
8. Nikolic N, Chevalet C. Detecting past changes of effective population size. Evolutionary Applications. 2014; 7(6):663–681. Available from: http://dx.doi.org/10.1111/eva.12170. doi: 10.1111/eva.12170 PMID: 25067949
9. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011; 475:493–496. doi: 10.1038/nature10231 PMID: 21753753
10. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nature Genetics. 2014; 46:919–925. doi: 10.1038/ng.3015 PMID: 24952747
11. Sheehan S, Harris K, Song YS. Estimating Variable Effective Population Sizes from Multiple Genomes: A Sequentially Markov Conditional Sampling Distribution Approach. Genetics. 2013; 194(3):647–662. Available from: http://www.genetics.org/content/194/3/647.abstract. doi: 10.1534/genetics.112.149096 PMID: 23608192
12. MacLeod IM, Larkin DM, Lewin HA, Hayes BJ, Goddard ME. Inferring Demography from Runs of Homozygosity in Whole-Genome Sequence, with Correction for Sequence Errors. Molecular Biology and Evolution. 2013; 30(9):2209–2223. Available from: http://mbe.oxfordjournals.org/content/30/9/2209.abstract. doi: 10.1093/molbev/mst125 PMID: 23842528
13. Bhaskar A, Wang YXR, Song YS. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. Genome Research. 2015; Available from:

http://genome.cshlp.org/content/early/2015/01/05/gr.178756.114.abstract. doi: 10.1101/gr.178756.114 PMID: 25564017

14. McVean GAT, Cardin NJ. Approximating the coalescent with recombination. Philosophical Transactions of the Royal Society B: Biological Sciences. 2005; 360(1459):1387–1393. Available from: http://rstb.royalsocietypublishing.org/content/360/1459/1387.abstract. doi: 10.1098/rstb.2005.1673

15. Marjoram P, Wall J. Fast "coalescent" simulation. BMC Genetics. 2006; 7(1):16. Available from: http://www.biomedcentral.com/1471-2156/7/16. doi: 10.1186/1471-2156-7-16 PMID: 16539698

16. Hudson RR. Properties of a neutral allele model with intragenic recombination. Theoret Popul Biol. 1983; 23:183–201. doi: 10.1016/0040-5809(83)90013-8

17. Sørensen AC, Sørensen MK, Berg P. Inbreeding in Danish Dairy Cattle Breeds. Journal of Dairy Science. 2005; 88 (5):1865–1872. doi: 10.3168/jds.S0022-0302(05)72861-7 PMID: 15829680

18. Boichard D, Maignel L, Verrier E. Analyse généalogique des races bovines laitières françaises. INRA Prod Anim. 1996; 9 (5):323–335.

19. Leroy G, Mary-Huard T, Verrier E, Danvy S, Charvolin E, Danchin-Burge C. Methods to estimate effective population size using pedigree data: Examples in dog, sheep, cattle and horse. Genetics Selection Evolution. 2013; 45(1):1. Available from: http://www.gsejournal.org/content/45/1/1. doi: 10.1186/1297-9686-45-1

20. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian Computation in Population Genetics. Genetics. 2002; 162(4):2025–2035. Available from: http://www.genetics.org/content/162/4/2025.abstract. PMID: 12524368

21. Wollstein A, Lao O, Becker C, Brauer S, Trent RJ, Nürnberg P, et al. Demographic History of Oceania Inferred from Genome-wide Data. Current Biology. 2010; 20(22):1983–1992. Available from: http://www.sciencedirect.com/science/article/pii/S0960982210013436. doi: 10.1016/j.cub.2010.10.040 PMID: 21074440

22. Li S, Jakobsson M. Estimating demographic parameters from large-scale population genomic data using Approximate Bayesian Computation. BMC Genetics. 2012; 13(1):22. Available from: http://www.biomedcentral.com/1471-2156/13/22. doi: 10.1186/1471-2156-13-22 PMID: 22453034

23. Theunert C, Tang K, Lachmann M, Hu S, Stoneking M. Inferring the History of Population Size Change from Genome-Wide SNP Data. Molecular Biology and Evolution. 2012; 29(12):3653–3667. Available from: http://mbe.oxfordjournals.org/content/29/12/3653.abstract. doi: 10.1093/molbev/mss175 PMID: 22787284

24. Nadachowska-Brzyska K, Burri R, Olason PI, Kawakami T, Smeds L, Ellegren H. Demographic Divergence History of Pied Flycatcher and Collared Flycatcher Inferred from Whole-Genome Re-sequencing Data. PLoS Genet. 2013 11; 9(11):e1003942. Available from: http://dx.doi.org/10.1371/journal.pgen.1003942. doi: 10.1371/journal.pgen.1003942 PMID: 24244198

25. Veeramah KR, Woerner AE, Johnstone L, Gut I, Gut M, Marques-Bonet T, et al. Examining Phylogenetic Relationships Among Gibbon Genera Using Whole Genome Sequence Data Using an Approximate Bayesian Computation Approach. Genetics. 2015; 200(1):295–308. Available from: http://www.genetics.org/content/200/1/295.abstract. doi: 10.1534/genetics.115.174425 PMID: 25769979

26. Hayes BJ, Visscher PM, McPartlan HC, Goddard ME. Novel Multilocus Measure of Linkage Disequilibrium to Estimate Past Effective Population Size. Genome Research. 2003; 13(4):635–643. Available from: http://genome.cshlp.org/content/13/4/635.abstract. doi: 10.1101/gr.387103 PMID: 12654718

27. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLoS Genet. 2009 10; 5(10):e1000695. Available from: http://dx.doi.org/10.1371/journal.pgen.1000695. doi: 10.1371/journal.pgen.1000695 PMID: 19851460

28. Lukić S, Hey J. Demographic Inference Using Spectral Methods on SNP Data, with an Analysis of the Human Out-of-Africa Expansion. Genetics. 2012; 192(2):619–639. Available from: http://www.genetics.org/content/192/2/619.abstract. doi: 10.1534/genetics.112.141846 PMID: 22865734

29. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust Demographic Inference from Genomic and SNP Data. PLoS Genet. 2013 10; 9(10):e1003905. Available from: http://dx.doi.org/10.1371/journal.pgen.1003905. doi: 10.1371/journal.pgen.1003905 PMID: 24204310

30. Patin E, Siddle KJ, Laval G, Quach H, Harmant C, Becker N, et al. The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. Nature Communications. 2014; 5: art. 3163. Available from: http://www.documentation.ird.fr/hor/fdi:010061680. doi: 10.1038/ncomms4163 PMID: 24495941

31. Liu X, Fu YX. Exploring population size changes using SNP frequency spectra. Nature genetics. 2015; 47(5):555–559. doi: 10.1038/ng.3254 PMID: 25848749

32. Blum M, Françcois O. Non-linear regression models for Approximate Bayesian Computation. Statistics and Computing. 2010; 20(1):63–73. Available from: http://dx.doi.org/10.1007/s11222-009-9116-0. doi: 10.1007/s11222-009-9116-0

33. Blum MGB, Nunes MA, Prangle D, Sisson SA. A comparative review of dimension reduction methods in approximate Bayesian computation. Statistical Science. 2013; 28:189–208. doi: 10.1214/12-STS406

34. Felius M, Beerling ML, Buchanan DS, Theunissen B, Koolmees PA, Lenstra JA. On the history of cattle genetic resources. Diversity. 2014; 6(4):705–750. doi: 10.3390/d6040705

35. Groenen M, Archibald A, Uenishi H, Tuggle C, Takeuchi Y, Rothschild M, et al. Analyses of pig genomes provide insight into porcine demography and evolution. Nature. 2012; 491:393–398. doi: 10.1038/nature11622 PMID: 23151582

36. Zhao S, Zheng P, Dong S, Zhan X, Wu Q, Guo X, et al. Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. Nature Genetics. 2013; 45(1):67–71. doi: 10.1038/ng.2494 PMID: 23242367

37. Green RE, Braun EL, Armstrong J, Earl D, Nguyen N, Hickey G, et al. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. Science. 2014; 346 (6215). Available from: http://www.sciencemag.org/content/346/6215/1254449.abstract. doi: 10.1126/science.1254449

38. Hung CM, Shaner PJL, Zink RM, Liu WC, Chu TC, Huang WS, et al. Drastic population fluctuations explain the rapid extinction of the passenger pigeon. Proceedings of the National Academy of Sciences. 2014; 111(29):10636–10641. Available from: http://www.pnas.org/content/111/29/10636.abstract. doi: 10.1073/pnas.1401526111

39. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, et al. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. PLoS Genet. 2014 4; 10(4):e1004234. Available from: http://dx.doi.org/10.1371/journal.pgen.1004234. doi: 10.1371/journal.pgen.1004234 PMID: 24743097

40. Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brondum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014; 46:858–865. doi: 10.1038/ng.3034 PMID: 25017103

41. Consortium TBH. Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. Science. 2009; 324(5926):528–532. Available from: http://www.sciencemag.org/content/324/5926/528.abstract. doi: 10.1126/science.1167936

42. Zeder MA. Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. Proceedings of the National Academy of Sciences. 2008; 105(33):11597–11604. Available from: http://www.pnas.org/content/105/33/11597.abstract. doi: 10.1073/pnas.0801317105

43. Edwards CJ, Ginja C, Kantanen J, Pérez-Pardal L, Tresset A, Stock F, et al. Dual Origins of Dairy Cattle Farming—Evidence from a Comprehensive Survey of European Y-Chromosomal Variation. PLoS ONE. 2011 1; 6(1):e15922. Available from: http://dx.doi.org/10.1371/journal.pone.0015922. doi: 10.1371/journal.pone.0015922 PMID: 21253012

44. Gautier M, Laloë D, Moazami-Goudarzi K. Insights into the Genetic History of French Cattle from Dense SNP Data on 47 Worldwide Breeds. PLoS ONE. 2010 9; 5(9):e13038. Available from: http://dx.doi.org/10.1371/journal.pone.0013038. doi: 10.1371/journal.pone.0013038 PMID: 20927341

45. Felius M, Koolmees PA, Theunissen B, Consortium ECGD, Lenstra JA. On the Breeds of Cattle—Historic and Current Classifications. Diversity. 2011; 3(4):660–692. Available from: http://www.mdpi.com/1424-2818/3/4/660. doi: 10.3390/d3040660

46. Csilléry K, Blum MGB, Gaggiotti OE, Françcois O. Approximate Bayesian Computation (ABC) in practice. Trends in Ecology and Evolution. 2010; 25(7):410–418. Available from: http://www.sciencedirect.com/science/article/pii/S0169534710000662. doi: 10.1016/j.tree.2010.04.001 PMID: 20488578

47. McEvoy BP, Powell JE, Goddard ME, Visscher PM. Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. Genome Research. 2011; Available from: http://genome.cshlp.org/content/early/2011/04/25/gr.119636.110.abstract. doi: 10.1101/gr.119636.110 PMID: 21518737

48. Hill WG. Estimation of effective population size from data on linkage disequilibrium. Genetics Research. 1981; 38:209–216. Available from: http://journals.cambridge.org/article_S0016672300020553. doi: 10.1017/S0016672300020553

49. Gattepaille LM, Jakobsson M, Blum M. Inferring population size changes with sequence and SNP data: lessons from human bottlenecks. Heredity. 2013; 110:409–419. doi: 10.1038/hdy.2012.120 PMID: 23423148

50. Rogers AR. How Population Growth Affects Linkage Disequilibrium. Genetics. 2014; Available from: http://www.genetics.org/content/early/2014/06/04/genetics.114.166454.abstract. doi: 10.1534/genetics.114.166454

51. Corbin LJ, Liu AYH, Bishop SC, Woolliams JA. Estimation of historical effective population size using linkage disequilibria with marker data. Journal of Animal Breeding and Genetics. 2012; 129(4):257–270. Available from: http://dx.doi.org/10.1111/j.1439-0388.2012.01003.x. doi: 10.1111/j.1439-0388.2012.01003.x PMID: 22775258

52. Bhaskar A, Song YS. Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. Ann Statist. 2014 12; 42(6):2469–2493. Available from: http://dx.doi.org/10.1214/14-AOS1264. doi: 10.1214/14-AOS1264

53. Bunnefeld L, Frantz LAF, Lohse K. Inferring Bottlenecks from Genome-Wide Samples of Short Sequence Blocks. Genetics. 2015; 201(3):1157–1169. Available from: http://www.genetics.org/content/201/3/1157.abstract. doi: 10.1534/genetics.115.179861 PMID: 26341659

54. Harris K, Nielsen R. Inferring Demographic History from a Spectrum of Shared Haplotype Lengths. PLoS Genet. 2013 6; 9(6):e1003521. Available from: http://dx.doi.org/10.1371/journal.pgen.1003521. doi: 10.1371/journal.pgen.1003521 PMID: 23754952

55. Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J. SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data. PLoS ONE. 2012 7; 7(7):e37558. Available from: http://dx.doi.org/10.1371/journal.pone.0037558. doi: 10.1371/journal.pone.0037558 PMID: 22911679

56. Maruki T, Lynch M. Genome-Wide Estimation of Linkage Disequilibrium from Population-Level High-Throughput Sequencing Data. Genetics. 2014; Available from: http://www.genetics.org/content/early/2014/05/27/genetics.114.165514.abstract. doi: 10.1534/genetics.114.165514

57. Flury C, Tapio M, Sonstegard T, Drögemüller C, Leeb T, Simianer H, et al. Effective population size of an indigenous Swiss cattle breed estimated from linkage disequilibrium. Journal of Animal Breeding and Genetics. 2010; 127(5):339–347. Available from: http://dx.doi.org/10.1111/j.1439-0388.2010.00862.x. doi: 10.1111/j.1439-0388.2010.00862.x PMID: 20831557

58. Shin DH, Cho KH, Park KD, Lee HJ, Kim H. Accurate Estimation of Effective Population Size in the Korean Dairy Cattle Based on Linkage Disequilibrium Corrected by Genomic Relationship Matrix. Asian Australas J Anim Sci. 2013; 26(12):1672–1679. Available from: http://ajas.info/journal/view.php?number=4618. doi: 10.5713/ajas.2013.13320 PMID: 25049757

59. Kijas JW, Lenstra JA, Hayes B, Boitard S, Porto Neto LR, San Cristobal M, et al. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. PLoS Biol. 2012; 10(2):e1001258. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=22346734. doi: 10.1371/journal.pbio.1001258 PMID: 22346734

60. Chikhi L, Sousa VC, Luisi P, Goossens B, Beaumont MA. The Confounding Effects of Population Structure, Genetic Diversity and the Sampling Scheme on the Detection and Quantification of Population Size Changes. Genetics. 2010; 186(3):983–995. Available from: http://www.genetics.org/content/186/3/983.abstract. doi: 10.1534/genetics.110.118661 PMID: 20739713

61. Peter BM, Wegmann D, Excoffier L. Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. Molecular Ecology. 2010; 19(21):4648–4660. Available from: http://dx.doi.org/10.1111/j.1365-294X.2010.04783.x. doi: 10.1111/j.1365-294X.2010.04783.x PMID: 20735743

62. Heller R, Chikhi L, Siegismund HR. The Confounding Effect of Population Structure on Bayesian Skyline Plot Inferences of Demographic History. PLoS ONE. 2013 5; 8(5):e62992. Available from: http://dx.doi.org/10.1371/journal.pone.0062992. doi: 10.1371/journal.pone.0062992 PMID: 23667558

63. Mazet O, Rodríguez W, Chikhi L. Demographic inference using genetic data from a single individual: Separating population size variation from population structure. Theoretical Population Biology. 2015; 104:46–58. Available from: http://www.sciencedirect.com/science/article/pii/S0040580915000581. doi: 10.1016/j.tpb.2015.06.003 PMID: 26120083

64. Mazet O, Rodriguez W, Grusea S, Boitard S, Chikhi L. On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference. Heredity. 2015; PMID: 26647653

65. Cornuet JM, Santos F, Beaumont MA, Robert CP, Marin JM, Balding DJ, et al. Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. Bioinformatics. 2008; 24(23):2713–2719. Available from: http://bioinformatics.oxfordjournals.org/content/24/23/2713.abstract. doi: 10.1093/bioinformatics/btn514 PMID: 18842597

66. Sandor C, Li W, Coppieters W, Druet T, Charlier C, Georges M. Genetic Variants in REC8, RNF212, and PRDM9 Influence Male Recombination in Cattle. PLoS Genet. 2012 7; 8(7):e1002854. Available from: http://dx.doi.org/10.1371/journal.pgen.1002854. doi: 10.1371/journal.pgen.1002854 PMID: 22844258

67. Rogers AR, Huff CD. Linkage Disequilibrium Between Loci With Unknown Phase. Genetics. 2009; 182:839–844. Available from: http://dx.doi.org/10.1534/genetics.108.093153. doi: 10.1534/genetics.108.093153 PMID: 19433632

68. Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 2002; 18:337–338. PMID: 11847089

69. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. The American Journal of Human Genetics. 2011; 88(1):76–82. Available from: http://www.sciencedirect.com/science/article/pii/S0002929710005987. doi: 10.1016/j.ajhg.2010.11.011 PMID: 21167468

70. Csilléry K, Françcois O, Blum MGB. abc: an R package for approximate Bayesian computation (ABC). Methods in Ecology and Evolution. 2012; 3(3):475–479. Available from: http://dx.doi.org/10.1111/j.2041-210X.2011.00179.x. doi: 10.1111/j.2041-210X.2011.00179.x

71. Tavaré S, Zeitouni O. Lectures on probability theory and statistics. Springer Berlin Heidelberg; 2004.

72. Hoze C, Fouilloux MN, Venot E, Guillaume F, Dassonneville R, Fritz S, et al. High-density marker imputation accuracy in sixteen French cattle breeds. Genetics Selection Evolution. 2013; 45(1):33. Available from: http://www.gsejournal.org/content/45/1/33. doi: 10.1186/1297-9686-45-33

Appendix B

Validating the number of differences

Comparing MS results to theoretical distributions

Contents

1 Comparing MS with theoretical distributions

1.1 Case 1. Single Wright-Fisher model. No structured constant-size population.

The expected coalescence time for two individuals in a Wright-Fisher model is $2N$ generations. That's the reason why the time is scaled by a factor of $2N$. In MS, the time is scaled by a factor of $4N$ which means that the expected coalescence time of two individuals will be 0.5. In order to compare the data simulated by MS with an exponential distribution, we need to set $\lambda = 2$. The following commands (in MS and python) should produce the same kind of data.

```
MS command:
ms 2 200000 -T | grep "(" | cut -d \: -f 2 | cut -d , -f 1

python command (using scipy)
scipy.stats.expon.rvs(scale=0.5, size=200000)
```

We analyze the outputs in this two cases for 200000 values. Results are shown in figure 1 in table 1.1.

Command	Mean	Variance	Min value	Max value
MS-command	0.50104192324499919	0.25043256012112991	6e-06	7.016047
python command	0.50008269094296265	0.25128866290586221	1.8743120203292514e-06	5.5221374492935205

Table 1: Comparing the two outputs of T_2 values

And the results of a KS-test are in table 2

Command	KS statistic	p-value
MS-command	0.0019646458602110006	0.42292233212463043
python command	0.001239959345653352	0.91819832476794261

Table 2: K-S test of the output data against the exp(2) distribution function. The blue curve is the theoretical distribution

In order to be sure that the output values of MS behaves as an exponential with $\lambda = 2$ we do the following experiment 1000 times:

1. Simulate 200000 values with the MS-command used before

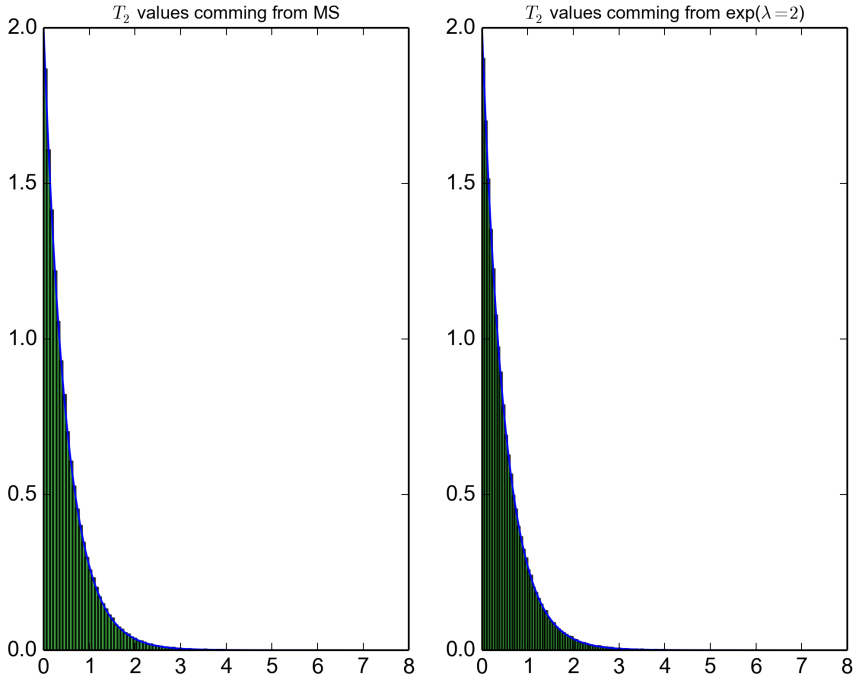


Figure 1: Histogram for T_2 values simulated with MS and exp(2). The theoretical function is in blue.

2. Simulate 200000 values with an exponential distribution ($\lambda = 2$)

3. Do a Kolmogorov-Smirnov test to compare the values with the theoretical exponential distribution (ie $f(x) = \lambda e^{-\lambda x}$

4. We reject the null hypothesis that the data comes from that distribution if the p-value is lower than 0.05

The number of times we rejected the null hypothesis was:

- 45 for data coming from MS ($4,5\%$ of reject)
- 48 for data coming from scipy.stats.expon.rvs(scale=0.5, size=200000) ($4,8\%$ of reject)

The experiment can be reproduced by using the "compare_w_MS.py" python module. The command is *KStest_MS_WFmodel(200000, 100, 0.05)*

1.2 Bottleneck with $\alpha = 2$ at time $T = 0.1$ case (SSPSC)

Now we consider a model where a population (with random mating) changes in size at time T in a factor of α.

1.2.1 Comparing T_2 values

Let's consider a model where the population decreased to a half of its size at time $T = 0.1$ (going forward in time). For simulating T_2 values under this model, the following MS-command can be used (note that MS counts time starting from the present, ie backward in the time):

```
MS 2 200000 -T -L -eN 0.1 2
```

We have to take into account that MS uses $T_{MS} = 4N$ generations and the distributions we have use $T = 2N$ generations.

$$P(T_2SSPSC < t) = P(2T_2MS < t)$$
$$P(T_2MS < t) = P(\frac{T_2SSPSC}{2} < t) = P(T_2SSPSC < 2t)$$

So, in order to compare the distribution function to the data coming from MS we have to use $F_{T_2SSPSC}(2t)$.

And in order to compare with the values produced by MS, the T parameter needs to be used as $2T$.

Figure 2 and Table 1.2.1 compare the outputs of the T_2 values produced by MS and by a method programming from the theoretical distribution.

Case	Mean	Variance	Min value	Max value
MS	0.90878557309999897	0.98751379447560694	1e-06	12.369535
SSPSC	0.90971717680963882	0.98378186798266054	4.202474792363943e-06	11.492613673166845

Table 3: Comparing the two outputs of T_2 values

Then, we do a KS-test for comparing both outputs with the theoretical distribution. Results are shown in table 4

Command	KS statistic	p-value
MS	0.0017264179181094574	0.59016399247324214
SSPSC	0.001200035111397102	0.93556672755384784

Table 4: K-S test of the output data against the theoretical SSPSC distribution function

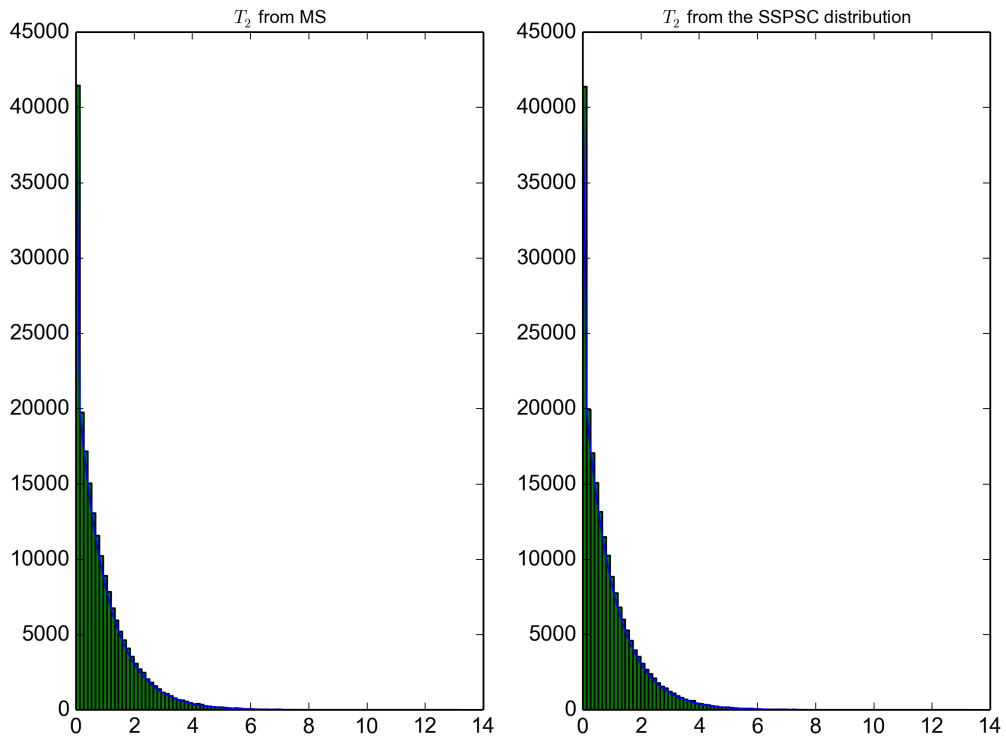


Figure 2: Histogram for T_2 values simulated with MS and SSPSC. The theoretical function is plotted in blue.

The KS-test was repeated 10000 times with $\alpha = 2$ and $T = 0.1$. We reject if the $p - value$ is lower than 0.05

The number of times we rejected the null hypothesis(*H0: The data actually comes from the theoretical distribution*) was:

- 524 for data coming from MS ($5,24\%$ of reject)
- 484 for data coming from SSPSC ($4,84\%$ of reject)

The experiment can be reproduced by using the "compare_MS_SSPSC.py" python script. The command used was

.compare_MS_SSPSC.py 2 0.1 200000 1

1.2.2 Comparing Number of SNPs

The *probability distribution function* of the number of mutation was derived in the article. For $t = 2N$ generations we have:

$$P(N_b = k) = \int_0^{+\infty} P(N_b = k|T_2^b = t) f_{T_2^b}(t) dt$$

When $t = 4N$ generations we have to change the parameters of the distribution in order to obtain the same values that MS produces.

In order to produce data (number of segregating sites) under this model, we can use the MS-command:

```
ms nobs nrep -t theta -eN T alpha
```

We used $alpha = 2, T = 0.1$ and $theta = 0.05$

The equivalent parameters for comparing our theoretical distribution to MS must be $2T$ for the time when changes occurred and $\theta/2$ for the mutation rate while keeping the same value of α.

We compare the results of the MS-command

```
./ms 2 20000 -t 0.05 -eN 0.1 2
```

with the output of our function for the corresponding parameters.

The results for one single experiment are in table 1.2.2. For the histogram see Figure 3. The results of a Chi-2 test for the same experiment are in table 6

Case	Mean	Variance	Min value	Max value
MS	0.91654999999999998	1.8696860974994656	0	12
SSPSC	0.90549999999999997	1.897169749999623	0	19

Table 5: Comparing the two outputs of $NumberofMutations$ values

Command	KS statistic	p-value
MS	15.799124639791151	0.10552889013725571
SSPSC	4.4964325296196161	0.95308785841375565

Table 6: Chi2 test of the output data against the theoretical SSPSC distribution function

We did 10000 independent repetitions of the experiment and we count the number of times we rejected the hypothesis that the data is coming from the theoretical distribution.

The number of times we rejected the null hypothesis was:

- 498 for data coming from MS ($4,98\%$ of reject)

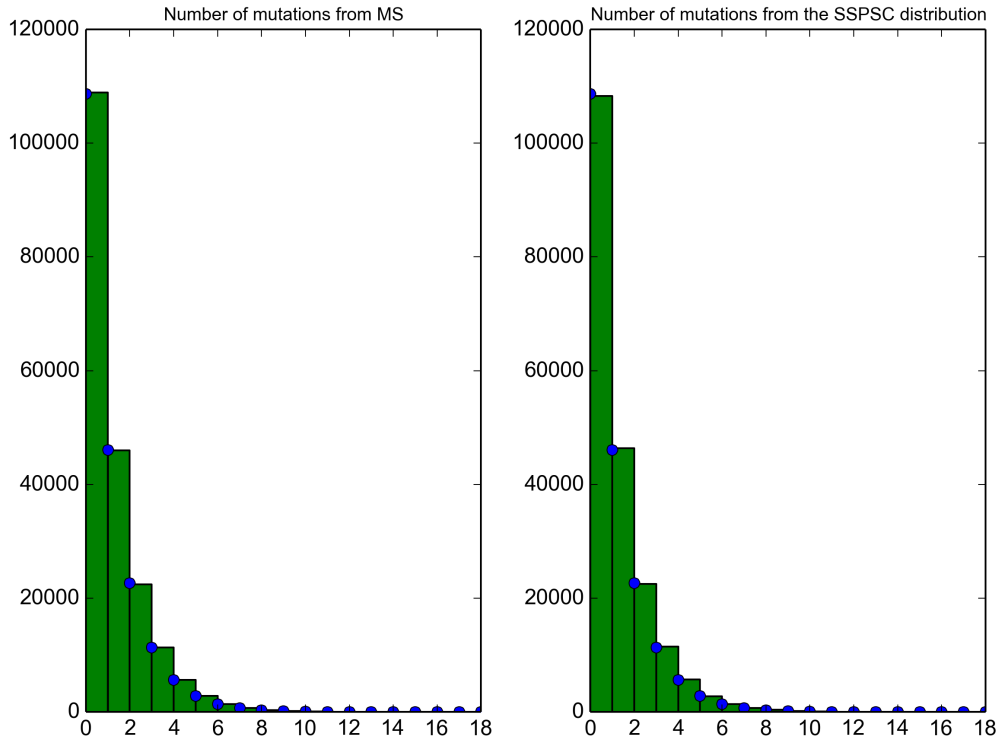


Figure 3: Histogram for $Number of Mutations$ values simulated with MS and SSPSC. The theoretical function is in blue.

- 486 for data coming from SSPSC ($4,86\%$ of reject)

The experiment can be reproduced by using the "compare_MS_SSPSC.py" python script. The command used was

.compare_MS_SSPSC.py 2 0.1 2000000 2 0.5

1.3 Structured population with $n = 9$ islands and $M = 0.1$ (StSI)

Now the model considered is the Symmetrical Island Model with 9 islands and migration rage of 0.1. We shall compare the values of T_2 as well as $Number\ of\ Mutations$ produced by MS with those produced by the method written from the theoretical distribution. Finally we do the corresponding KS and Chi2 tests.

1.3.1 Comparing the T_2 values

The MS command for simulating the values of coalescence times when two individuals are sampled from the same island in this population is

```
MS 2 200000 -T -L -I 9 2 0 0 0 0 0 0 0 0 0.1
```

We have to take into account that MS uses $T_{MS} = 4N$ generations and the distribution we have uses $T = 2N$ generations. So, in order to compare the distribution function to the data coming from MS we have to use $F_{T_2 StSI}(2t)$

The Figure 4 and the table 7 show the results of 200000 independent values simulated using MS and StSI distribution function.

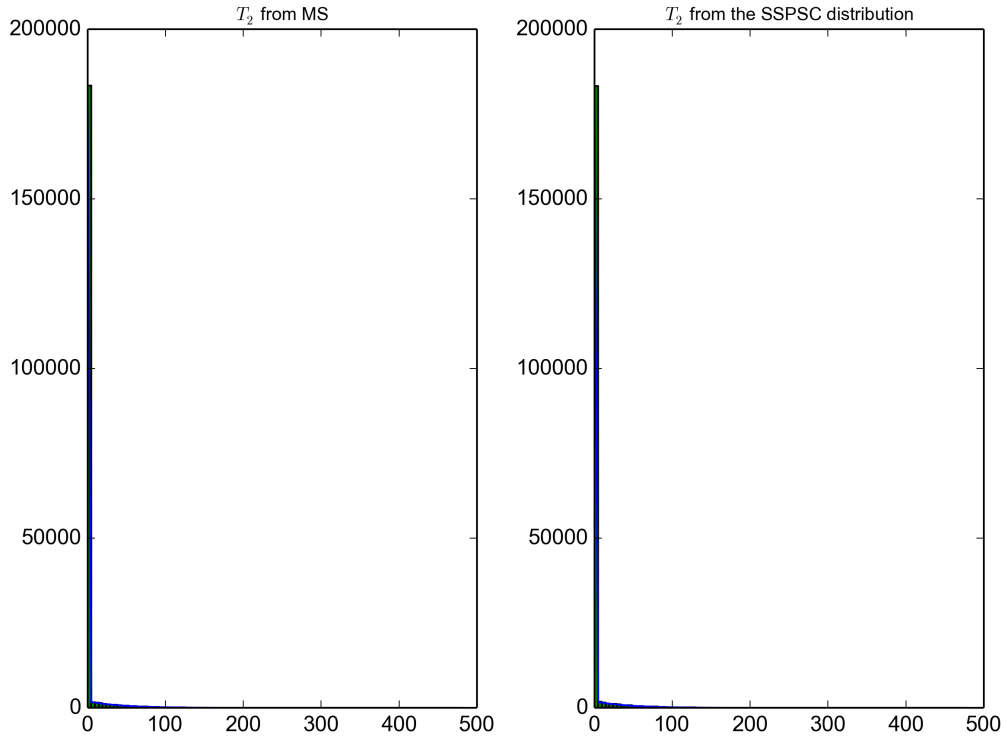


Figure 4: Histogram for T_2 values simulated with MS and StSI. The theoretical function is in blue.

And the results of a KS-test are in table 8

As before, we did a KS test 10000 times with $n = 9$ and $M = 0.1$

The number of times we rejectet the null hypothesis (*H0: The values come from the theoretical distribution of T_2 for this model*) was:

Case	Mean	Variance	Min value	Max value
MS	4.5176277987000466	347.11642120835103	7e-06	441.000519
StSI	4.5722299125887247	346.77477764513173	4.7609152318513286e-06	484.13932749856457

Table 7: Comparing the two outputs of T_2 values

Command	KS statistic	p-value
MS	0.0013742768784799075	0.84438710150019836
StSI	0.0014041958586083481	0.82521917262152

Table 8: K-S test of the output data against the theoretical StSI distribution function

- 499 for data coming from MS ($4,99\%$ of reject)
- 503 for data coming from StSI ($5,03\%$ of reject)

The experiment can be reproduced by using the "compare_MS_StSI.py" python script. The command used was

.compare_MS_StSI.py 9 0.1 200000 1

1.3.2 Comparing Number of SNPs

Now we compare the $Number\ of\ Mutations$ variable for both cases.

The Figure 5 and the table 9 show the results of 200000 independent values simulated by MS and StSI distribution function.

Case	Mean	Variance	Min value	Max value
MS	4.5271249999999998	349.74466423523086	0	437
StSI	4.4704100000000002	346.97783443283259	0	441

Table 9: Comparing the two outputs of $NumberofMutations$ values

And the results of a Chi2-test are in table 10

We did 10000 independent repetitions of the experiment and we count the number of times we rejected the hypothesis that the data is coming from the theoretical distribution.

The number of times we rejected the null hypothesis was:

- 526 for data comming from MS ($5,26\%$ of reject)
- 522 for data comming from SSPSC ($5,22\%$ of reject)

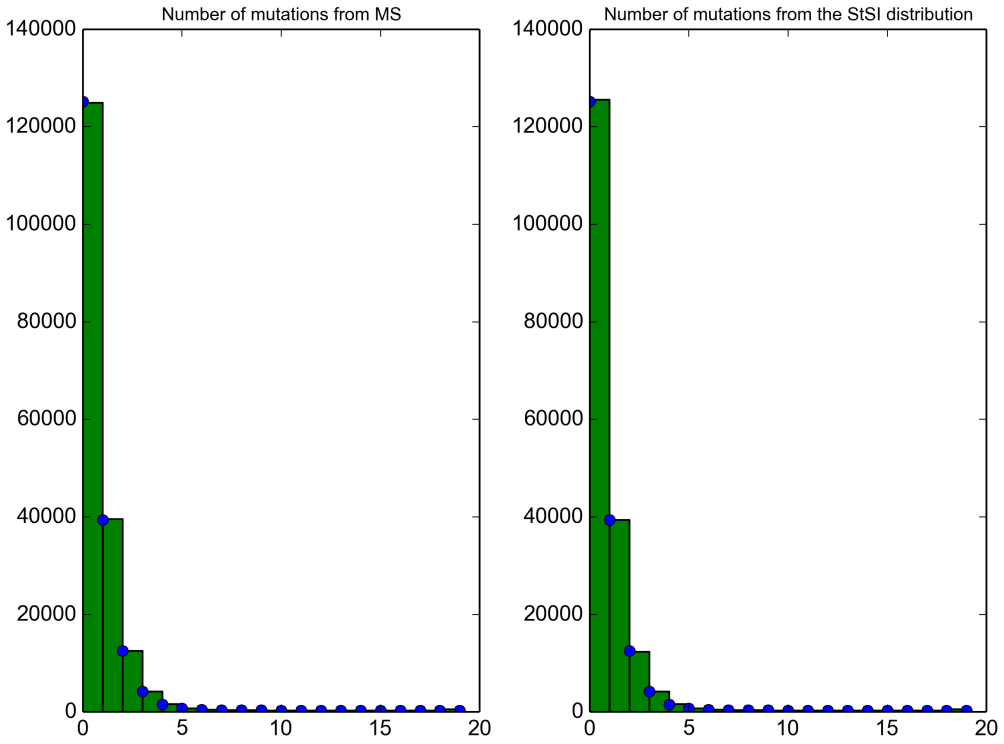


Figure 5: Histogram for $Number of Mutations$ values simulated with MS and StSI. The theoretical function is in blue.

The experiment can be reproduced by using the "compare_MS_SSPSC.py" python script. The command used was

.compare_MS_StSI.py 9 0.1 2000000 2 0.5

Command	KS statistic	p-value
MS	193.57103521370334	0.43445621518516619
StSI	186.83623084924969	0.61144954519283146

Table 10: Chi2 test of the $NumberofMutations$ output data against the theoretical StSI distribution function

Appendix C

Validating a python implementation of the NIMC

Validating the implementation of the NIMC

1 Some validations of the NIMC implementation

The *NIMC* module is a python implementation of the N-Island Markov Chain model. In order to validate this implementation of the *NIMC*, some comparison are done between the cumulative distribution function (*cdf*) as well as the probability density function (*pdf*) of T_2^s and T_2^d (the coalescence times of two genes sampled in the *same* population or in *different* populations) implemented in the *NIMC* class, with the empirical distribution of the values simulated with the **ms** software under equivalent scenarios. It is important to note that the time in *ms* is scaled to $4N_0$ while all the theoretical computations used in the *NIMC* use a time scale of $2N_0$. This implies that, in order to do the comparisons, it is necessary to multiply the output times comming from ms by 2 and the times when demographic events occur for the corresponding ms commands should be divided by 2 (i.e. if the gene flow in the NIMC model changes at time $T = 1$, it should be used $T = 0.5$ when traslating this into the corresponding ms command).

```
In [1]: import tester_NIMC_T2
        from tester_NIMC_T2 import tester
        %matplotlib inline
        t = tester()
```

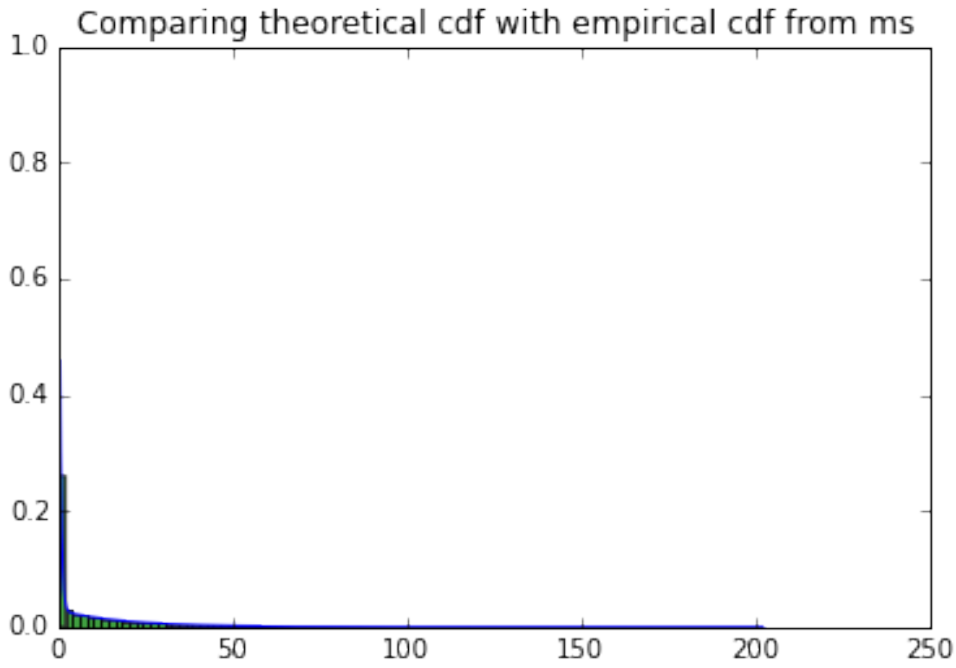
1.1 Testing T_2^s and T_2^d under an n-island model

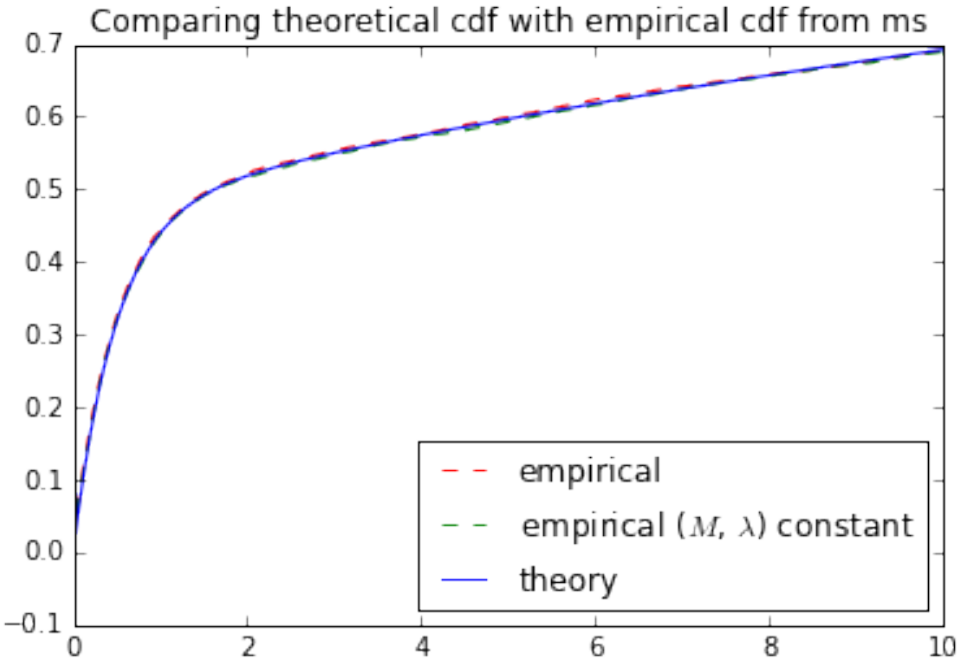
```
In [2]: n = 10
        T_list = [0]
        M_list = [1]
        lambda_list = [1]
        n_obs = 10000
        number_of_tests = 100

        t.do_full_comparison(n, T_list, M_list, lambda_list, n_obs, number_of_tests)
```

```
Sampling in the same island:
------------------------------

The corresponding ms-command is:
./utils/ms 2 10000 -T -L -I 10 2 0 0 0 0 0 0 0 0 0 1.0
```



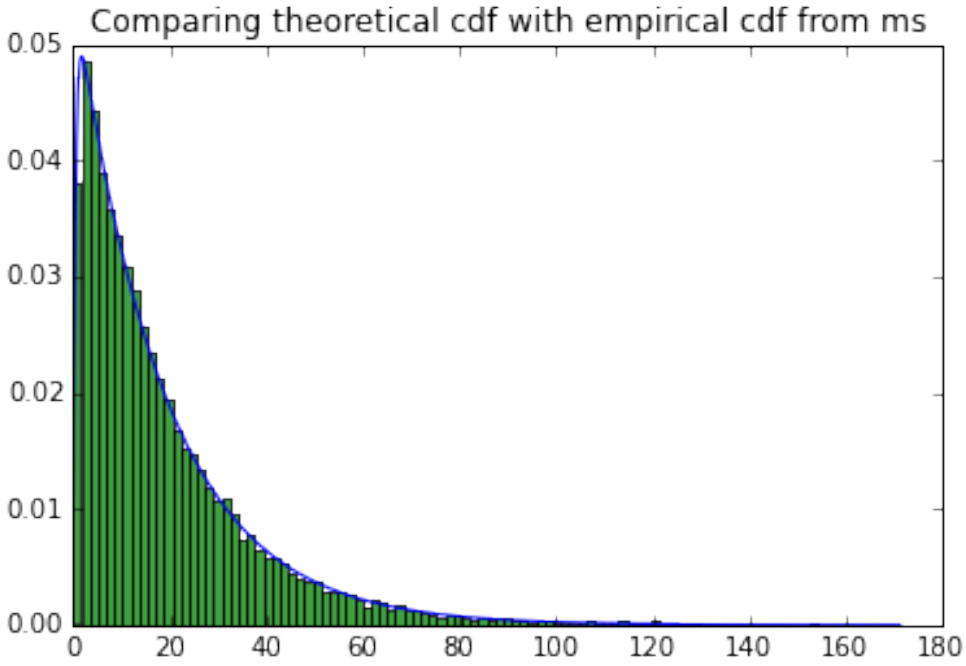


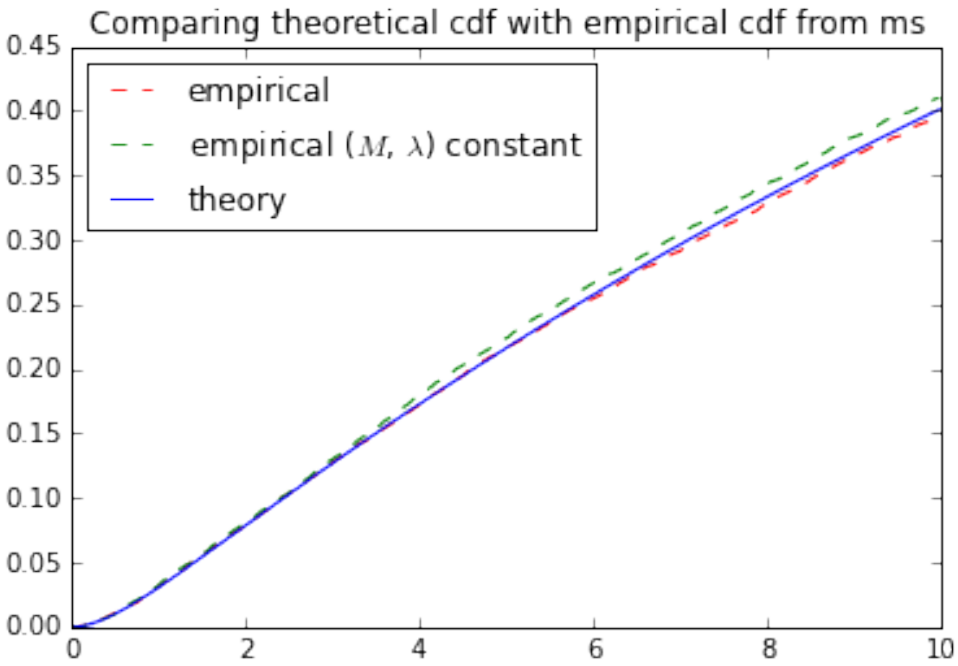

```
the p-value of one single ks-test: 0.509534615757
Doing 100 ks-tests ...
```

```
Number of rejections with alpha=0.05: 5

Sampling in different islands:
------------------------------

The corresponding ms-command is:
./utils/ms 2 10000 -T -L -I 10 1 1 0 0 0 0 0 0 0 0 1.0
```






```
the p-value of one single ks-test: 0.474966924017
Doing 100 ks-tests ...
Number of rejections with alpha=0.05: 4
```

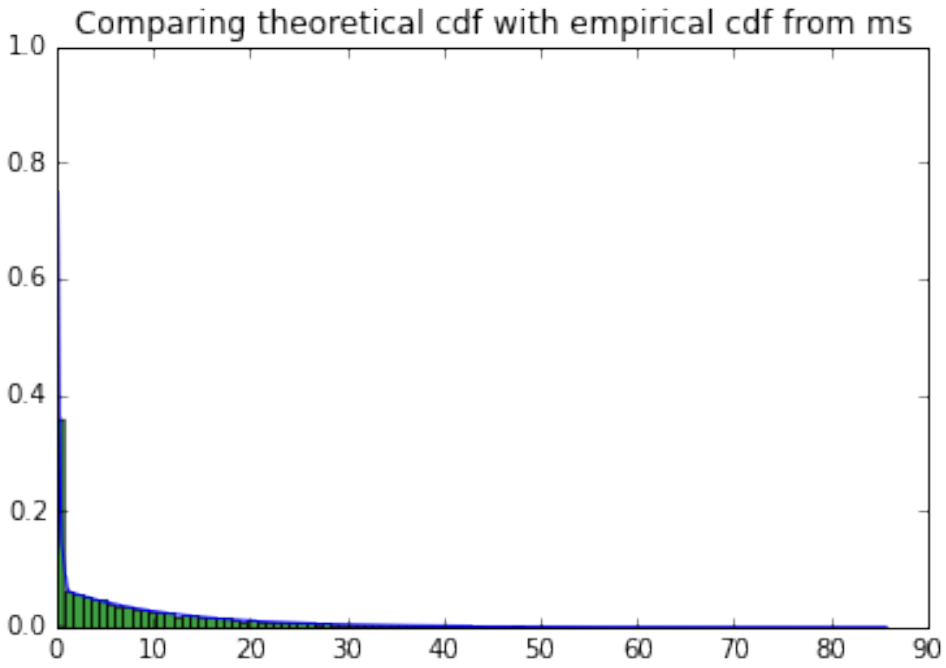
1.2 Adding some gene flow changes

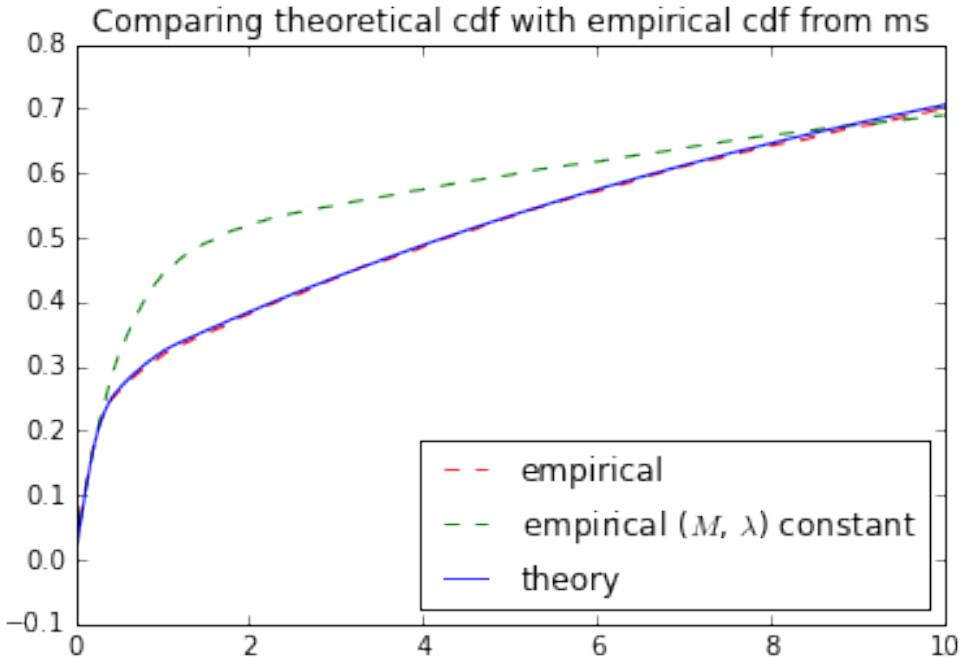
```
In [3]: n = 10
        T_list = [0, 0.2, 0.5, 1]
        M_list = [1, 5, 0.1, 10]
        lambda_list = [1, 1, 1, 1]
        n_obs = 10000
        number_of_tests = 100

        t.do_full_comparison(n, T_list, M_list, lambda_list, n_obs, number_of_tests)
```

```
Sampling in the same island:
----------------------------

The corresponding ms-command is:
./utils/ms 2 10000 -T -L -I 10 2 0 0 0 0 0 0 0 0 0 0 1.0 -eM 0.1 5.0 -eN 0.1 1.0 -eM 0.25 0.1 -eN
    0.25 1.0 -eM 0.5 10.0 -eN 0.5 1.0
```

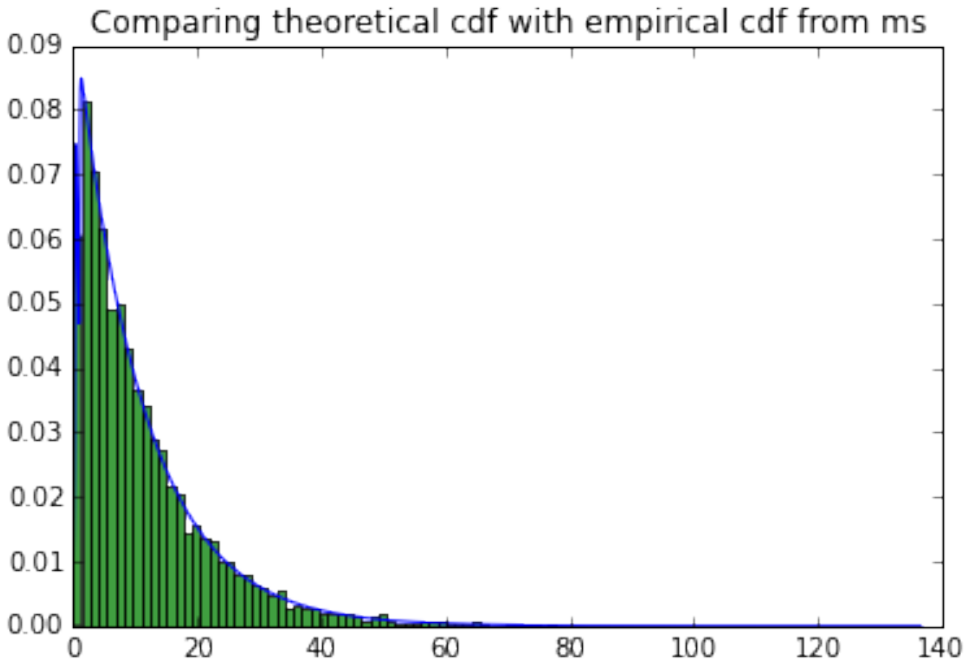


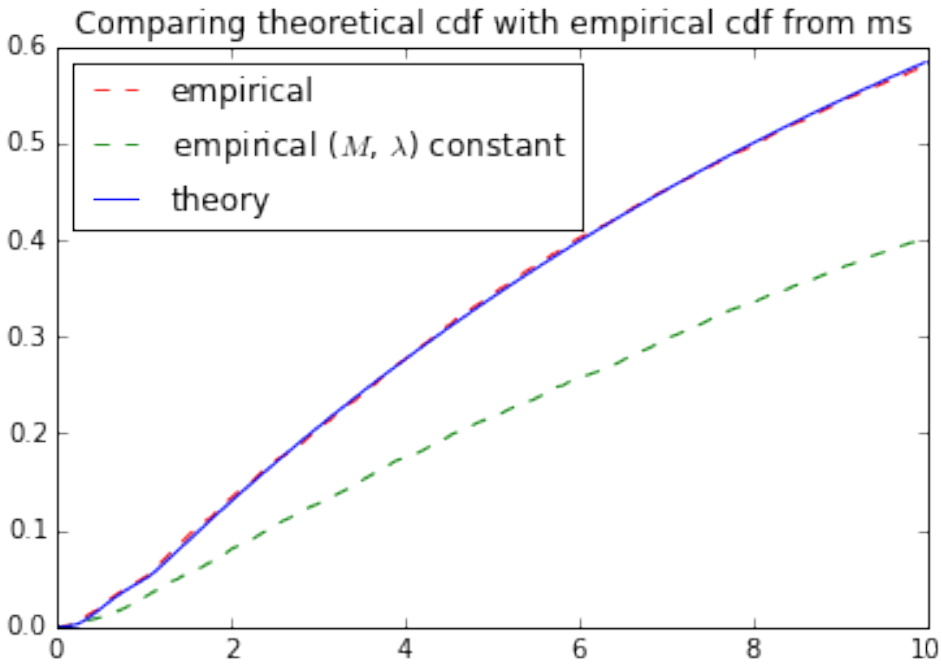



```
the p-value of one single ks-test: 0.40911489094
Doing 100 ks-tests {\ldots}
Number of rejections with alpha=0.05: 3
```

```
Sampling in different islands:
------------------------------

The corresponding ms-command is:
./utils/ms 2 10000 -T -L -I 10 1 1 0 0 0 0 0 0 0 0 1.0 -eM 0.1 5.0 -eN 0.1 1.0 -eM 0.25 0.1 -eN
    0.25 1.0 -eM 0.5 10.0 -eN 0.5 1.0
```






```
the p-value of one single ks-test: 0.96474837359
Doing 100 ks-tests ...
Number of rejections with alpha=0.05: 6
```

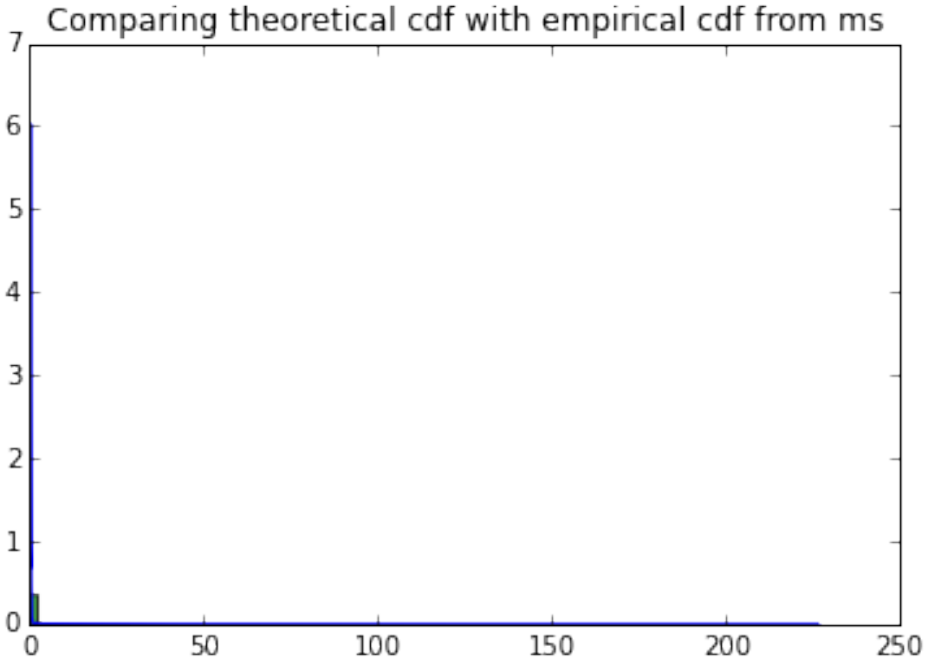
1.3 Considering just population size changes with constant gene flow

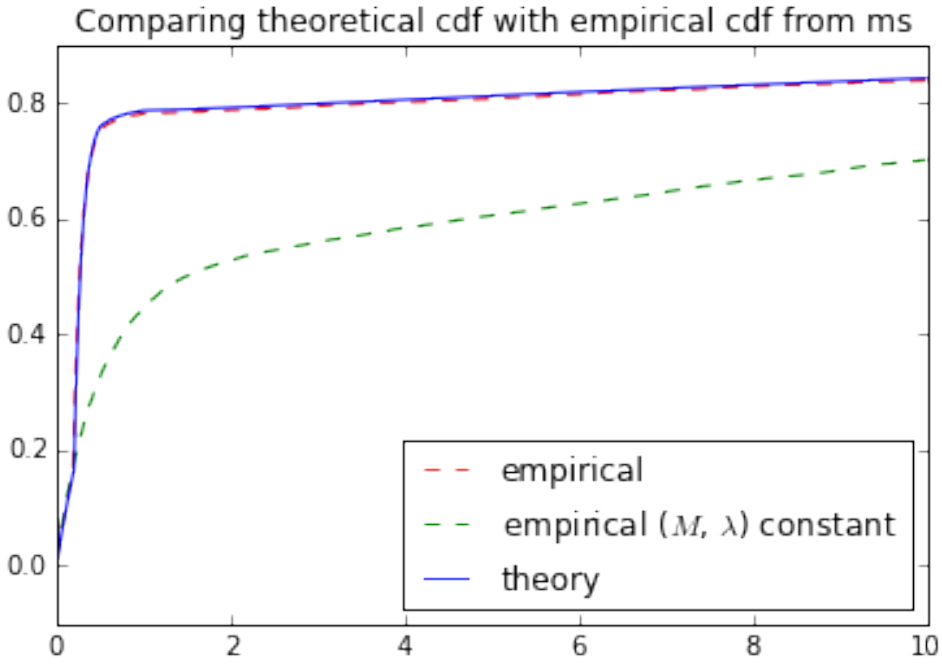
```
In [4]: n = 10
        T_list = [0, 0.2, 0.5, 1]
        M_list = [1, 1, 1, 1]
        lambda_list = [1, 10, 5, 0.5]
        n_obs = 10000
        number_of_tests = 100

        t.do_full_comparison(n, T_list, M_list, lambda_list, n_obs, number_of_tests)
```

```
Sampling in the same island:
------------------------------

The corresponding ms-command is:
./utils/ms 2 10000 -T -L -I 10 2 0 0 0 0 0 0 0 0 0 0 1.0 -eM 0.1 1.0 -eN 0.1 0.1 -eM 0.25 1.0 -eN
    0.25 0.2 -eM 0.5 1.0 -eN 0.5 2.0
```

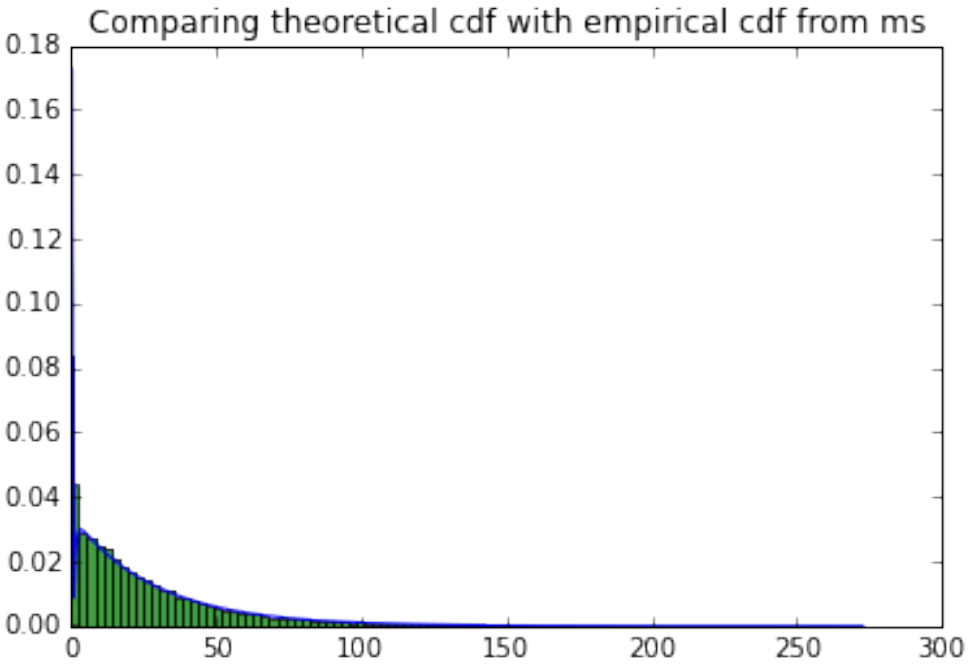


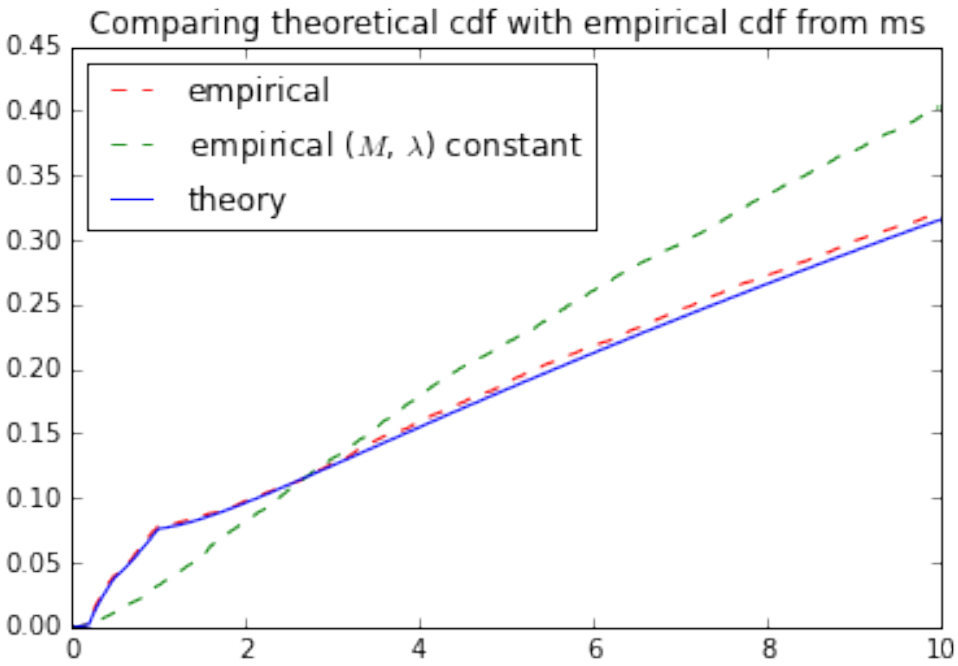



```
the p-value of one single ks-test: 0.955687502168
Doing 100 ks-tests {\ldots}
Number of rejections with alpha=0.05: 6
```

```
Sampling in different islands:
------------------------------

The corresponding ms-command is:
./utils/ms 2 10000 -T -L -I 10 1 1 0 0 0 0 0 0 0 0 1.0 -eM 0.1 1.0 -eN 0.1 0.1 -eM 0.25 1.0 -eN
    0.25 0.2 -eM 0.5 1.0 -eN 0.5 2.0
```






```
the p-value of one single ks-test: 0.217524751657
Doing 100 ks-tests ...
Number of rejections with alpha=0.05: 8
```

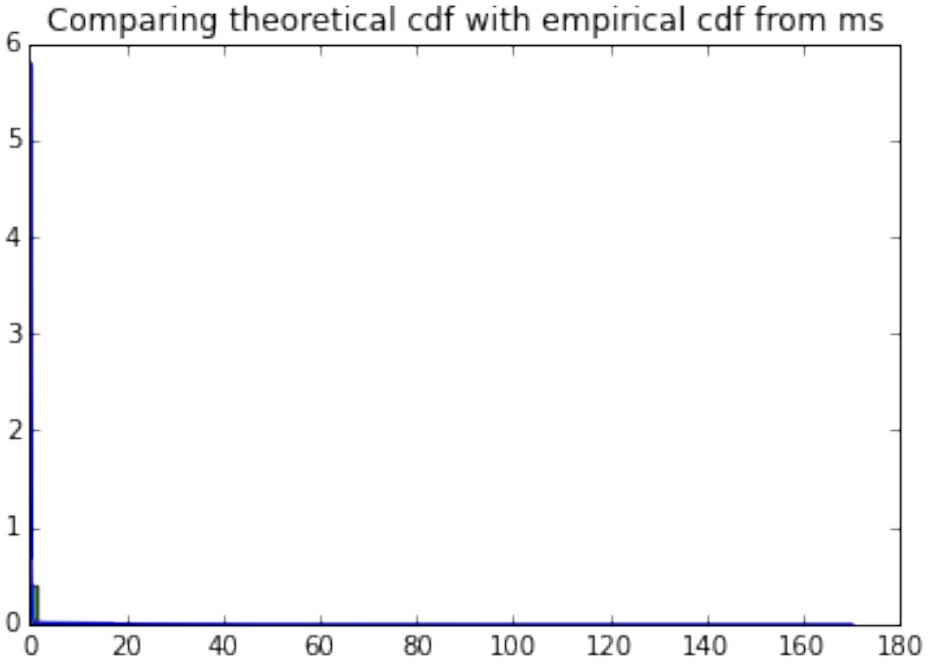
1.4 Now changing both, gene flow and population size

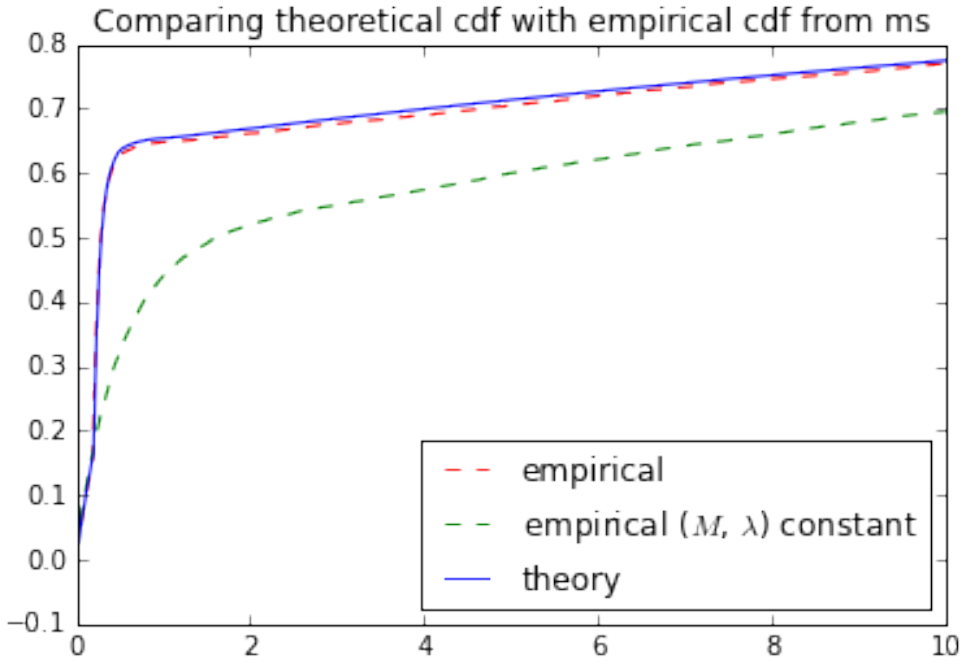
```
In [5]: n = 10
        T_list = [0, 0.2, 0.5, 1]
        M_list = [1, 5, 0.1, 10]
        lambda_list = [1, 10, 5, 0.5]
        n_obs = 10000
        number_of_tests = 100

        t.do_full_comparison(n, T_list, M_list, lambda_list, n_obs, number_of_tests)
```

```
Sampling in the same island:
------------------------------

The corresponding ms-command is:
./utils/ms 2 10000 -T -L -I 10 2 0 0 0 0 0 0 0 0 0 0 1.0 -eM 0.1 5.0 -eN 0.1 0.1 -eM 0.25 0.1 -eN
    0.25 0.2 -eM 0.5 10.0 -eN 0.5 2.0
```

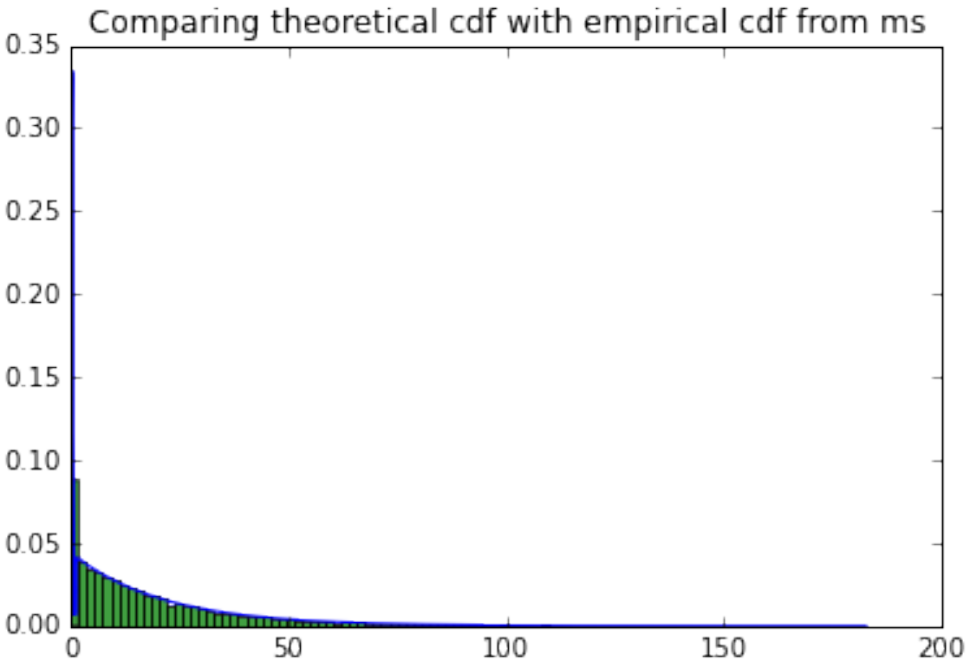


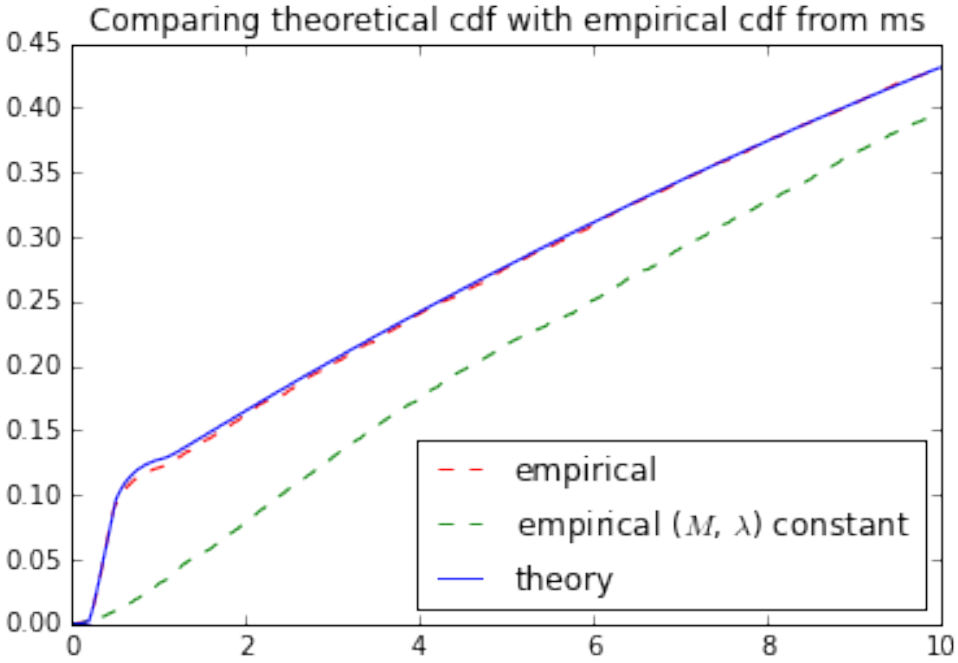



```
the p-value of one single ks-test: 0.411042128269
Doing 100 ks-tests {\ldots}
Number of rejections with alpha=0.05: 6
```

```
Sampling in different islands:
------------------------------

The corresponding ms-command is:
./utils/ms 2 10000 -T -L -I 10 1 1 0 0 0 0 0 0 0 0 1.0 -eM 0.1 5.0 -eN 0.1 0.1 -eM 0.25 0.1 -eN
    0.25 0.2 -eM 0.5 10.0 -eN 0.5 2.0
```






```
the p-value of one single ks-test: 0.213663349949
Doing 100 ks-tests ...
Number of rejections with alpha=0.05: 6
```

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. Automatic Control, IEEE Transactions on, 19(6):716–723.

Bahlo, M. and Griffiths, R. C. (2001). Coalescence time for two genes from a subdivided population. Journal of Mathematical Biology, 43(5):397–410.

Barton, N. H., Depaulis, F., and Etheridge, A. M. (2002). Neutral evolution in spatially continuous populations. Theoretical Population Biology, 61(1):31 – 48.

Barton, N. H. and Wilson, I. (1995). Genealogies and geography. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 349(1327):49–59.

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

Beaumont, M. A. (2010). Approximate bayesian computation in evolution and ecology. Annual review of ecology, evolution, and systematics, 41:379–406.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate bayesian computation in population genetics. Genetics, 162(4):2025–2035.

Boitard, S., Rodríguez, W., Jay, F., Mona, S., and Austerlitz, F. (2016). Inferring population size history from large samples of genome-wide molecular data - an approximate bayesian computation approach. PLoS Genet, 12(3):1–36.

Broquet, T., Angelone, S., Jaquiery, J., Joly, P., Lena, J.-P., Lengagne, T., Plenet, S., Luquet, E., and Perrin, N. (2010). Genetic bottlenecks driven by population disconnection. Conservation Biology, 24(6):1596–1605.

Cahill, J. A., Green, R. E., Fulton, T. L., Stiller, M., Jay, F., Ovsyanikov, N., Salamzade, R., St. John, J., Stirling, I., Slatkin, M., and Shapiro, B. (2013). Genomic evidence for island population conversion resolves conflicting theories of polar bear evolution. PLoS Genet, 9(3):1–8.

Cannings, C. (1974). The latent roots of certain markov chains arising in genetics: A new approach, i. haploid models. Advances in Applied Probability, 6(2):260–290.

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. Nat Rev Genet, 10(3):195–205.

Charlesworth, B., Charlesworth, D., and Barton, N. H. (2003). The effects of genetic and geographic structure on neutral variation. Annual Review of Ecology, Evolution, and Systematics, pages 99–125.

Chen, C., Durand, E., Forbes, F., and François, O. (2007). Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. Molecular Ecology Notes, 7(5):747–756.

Chikhi, L., Sousa, V. C., Luisi, P., Goossens, B., and Beaumont, M. A. (2010). The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. Genetics, 186(3):983–995.

Corander, J., Waldmann, P., Marttinen, P., and Sillanpää, M. J. (2004). Baps 2: enhanced possibilities for the analysis of genetic population structure. Bioinformatics, 20(15):2363–2369.

Cornuet, J.-M., Santos, F., Beaumont, M. A., Robert, C. P., Marin, J.-M., Balding, D. J., Guillemaud, T., and Estoup, A. (2008). Inferring population history with diy abc: a user-friendly approach to approximate bayesian computation. Bioinformatics, 24(23):2713–2719.

Crow, J. F. (1988). Eighty years ago: the beginnings of population genetics. Genetics, 119(3):473.

Csilléry, K., Blum, M. G., Gaggiotti, O. E., and François, O. (2010). Approximate bayesian computation (abc) in practice. Trends in Ecology & Evolution, 25(7):410 – 418.

Currat, M., Excoffier, L., Maddison, W., Otto, S. P., Ray, N., Whitlock, M. C., and Yeaman, S. (2006). Comment on "ongoing adaptive evolution of aspm, a brain size determinant in homo sapiens" and "microcephalin, a gene regulating brain size, continues to evolve adaptively in humans". Science, 313(5784):172a–172a.

Delmas, J.-F. and Jourdain, B. (2006). Modèles aléatoires. Springer.

Donnelly, P. and Tavaré, S. (1995). Coalescents and genealogical structure under neutrality. Annual Review of Genetics, 29(1):401–421. PMID: 8825481.

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. Molecular Biology and Evolution, 22(5):1185–1192.

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

Ewens, W. J. (2012). Mathematical Population Genetics 1: Theoretical Introduction, volume 27. Springer Science & Business Media.

Felsenstein, J. (1981). Evolutionary trees from dna sequences: A maximum likelihood approach. Journal of Molecular Evolution, 17(6):368–376.

Fisher, R. A. (1930). The genetical theory of natural selection: a complete variorum edition. Oxford University Press.

Girod, C., Vitalis, R., Leblois, R., and Fréville, H. (2011). Inferring population decline and expansion from microsatellite data: A simulation-based evaluation of the msvar method. Genetics, 188(1):165–179.

Goldstein, D. B. and Chikhi, L. (2002). Human migrations and population structure: what we know and why it matters. Annual Review of Genomics and Human Genetics, 3(1):129–152.

Goossens, B., Chikhi, L., Ancrenaz, M., Lackman-Ancrenaz, I., Andau, P., Bruford, M. W., et al. (2006). Genetic signature of anthropogenic population collapse in orang-utans. PLoS Biology, 4(2):285.

Green, R. E., Braun, E. L., Armstrong, J., Earl, D., Nguyen, N., Hickey, G., Vandewege, M. W., St. John, J. A., Capella-Gutiérrez, S., Castoe, T. A., Kern, C., Fujita, M. K., Opazo, J. C., Jurka, J., Kojima, K. K., Caballero, J., Hubley, R. M., Smit, A. F., Platt, R. N., Lavoie, C. A., Ramakodi, M. P., Finger, J. W., Suh, A., Isberg, S. R., Miles, L., Chong, A. Y., Jaratlerdsiri, W., Gongora, J., Moran, C., Iriarte, A., McCormack, J., Burgess, S. C., Edwards, S. V., Lyons, E., Williams, C., Breen, M., Howard, J. T., Gresham, C. R., Peterson, D. G., Schmitz, J., Pollock, D. D., Haussler, D., Triplett, E. W., Zhang, G., Irie, N., Jarvis, E. D., Brochu, C. A., Schmidt, C. J., McCarthy, F. M., Faircloth, B. C., Hoffmann, F. G., Glenn, T. C., Gabaldón, T., Paten, B., and Ray, D. A. (2014). Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. Science, 346(6215).

Griffiths, R. (1981). The number of heterozygous loci between two randomly chosen completely linked sequences of loci in two subdivided population models. Journal of Mathematical Biology, 12(2):251–261.

Griffiths, R. and Tavaré, S. (1994). Simulating probability distributions in the coalescent. Theoretical Population Biology, 46(2):131–159.

Griffiths, R. C. and Tavaré, S. (1994). Sampling theory for neutral alleles in a varying environment. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 344(1310):403–410.

Groenen, M., Archibald, A., Uenishi, H., Tuggle, C., Takeuchi, Y., Rothschild, M., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H., Li, S., Larkin, D., Kim, H., Frantz, L., Caccamo, M., Ahn, H., Aken, B., Anselmo, A., Anthon, C., Auvil, L., Badaoui, B., Beattie, C., Bendixen, C., Berman, D., Blecha, F., Blomberg, J., Bolund, L., Bosse, M., Botti, S., and Bujie, Z. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. Nature, 491:393–398.

Guillot, G., Mortier, F., and Estoup, A. (2005). Geneland: a computer package for landscape genetics. Molecular Ecology Notes, 5(3):712–715.

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 Genet, 5(10):e1000695.

Hallatschek, O. and Fisher, D. S. (2014). Acceleration of evolutionary spread by long-range dispersal. Proceedings of the National Academy of Sciences, 111(46):E4911–E4919.

Harding, R. M. and McVean, G. (2004). A structured ancestral population for the evolution of modern humans. Current opinion in genetics & development, 14(6):667–674.

Hardy, G. H. (1908). Mendelian proportions in a mixed population. Science, 28(706):pp. 49–50.

Harpending, H. and Rogers, A. (2000). Genetic perspectives on human origins and differentiation. Annual Review of Genomics and Human Genetics, 1(1):361–385.

Heller, R., Chikhi, L., and Siegismund, H. R. (2013). The confounding effect of population structure on bayesian skyline plot inferences of demographic history. PLoS ONE, 8(5):e62992.

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

Hey, J. and Machado, C. A. (2003). The study of structured populations - new hope for a difficult and divided science. Nat Rev Genet, 4(7):535–543.

Hobolth, A., Andersen, L. N., and Mailund, T. (2011). On computing the coalescence time density in an isolation-with-migration model with few samples. Genetics, 187(4):1241–1243.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology, 23(2):183 – 201.

Hudson, R. R. (2002). Generating samples under a wright–fisher neutral model of genetic variation. Bioinformatics, 18(2):337–338.

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

Hung, C.-M., Shaner, P.-J. L., Zink, R. M., Liu, W.-C., Chu, T.-C., Huang, W.-S., and Li, S.-H. (2014). Drastic population fluctuations explain the rapid extinction of the passenger pigeon. Proceedings of the National Academy of Sciences, 111(29):10636–10641.

Ian W. Saunders, Simon Tavaré, G. A. W. (1984). On the genealogy of nested subsamples from a haploid population. Advances in Applied Probability, 16(3):471–491.

Jones, E., Oliphant, T., Peterson, P., et al. (2001). Scipy: Open source scientific tools for python. [Online; accessed 2014-11-18].

Kimura, M. and Weiss, G. H. (1964). The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics, 49(4):561–76.

Kingman, J. (1982a). The coalescent. Stochastic Processes and their Applications, 13(3):235 – 248.

Kingman, J. F. C. (1982b). Exchangeability and the evolution of large populations. In Koch, G. and Spizzichino, F., editors, Exchangeability in Probability and Statistics, pages 97–112. North-Holland, Amsterdam.

Kingman, J. F. C. (1982c). On the genealogy of large populations. Journal of Applied Probability, 19:27–43.

Klein, J. P. and Moeschberger, M. L. (2003). Survival analysis: techniques for censored and truncated data. Springer Science & Business Media.

Leblois, R., Estoup, A., and Streiff, R. (2006). Genetics of recent habitat contraction and reduction in population size: does isolation by distance matter? Molecular Ecology, 15(12):3601–3615.

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

Liu, X. and Fu, Y.-X. (2015). Exploring population size changes using snp frequency spectra. Nature genetics.

MacLeod, I. M., Larkin, D. M., Lewin, H. A., Hayes, B. J., and Goddard, M. E. (2013). Inferring demography from runs of homozygosity in whole-genome sequence, with correction for sequence errors. Molecular Biology and Evolution, 30(9):2209–2223.

Marjoram, P. and Wall, J. D. (2006). Fast "coalescent" simulation. BMC Genetics, 7(1):1–9.

Mazet, O., Rodriguez, W., Grusea, S., Boitard, S., and Chikhi, L. (2015a). On the importance of being structured: instantaneous coalescence rates and human evolution[mdash]lessons for ancestral population size inference[quest]. Heredity, pages –.

Mazet, O., Rodríguez, W., and Chikhi, L. (2015b). Demographic inference using genetic data from a single individual: Separating population size variation from population structure. Theoretical Population Biology, 104:46 – 58.

McManus, K. F., Kelley, J. L., Song, S., Veeramah, K., Woerner, A. E., Stevison, L. S., Ryder, O. A., Kidd, J. M., Wall, J. D., Bustamante, C. D., and Hammer, M. (2015). Inference of gorilla demographic and selective history from whole genome sequence data. Molecular Biology and Evolution, pages 600–612.

McVean, G. A. and Cardin, N. J. (2005a). Approximating the coalescent with recombination. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 360(1459):1387–1393.

McVean, G. A. T. and Cardin, N. J. (2005b). Approximating the coalescent with recombination. Philos. Trans. R. Soc. Lond., B, Biol. Sci., 360(1459):1387–93.

Moler, C. and Loan, C. V. (2003). Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Review, 45(1):3–49.

Moran, P. A. P. (1958). Random processes in genetics. Mathematical Proceedings of the Cambridge Philosophical Society, 54:60–71.

Nagylaki, T. (1980). The strong-migration limit in geographically structured populations. Journal of Mathematical Biology, 9(2):101–114.

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

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. The computer journal, 7(4):308–313.

Nielsen, R. and Beaumont, M. A. (2009). Statistical inferences in phylogeography. Molecular Ecology, 18(6):1034–1047.

Nielsen, R. and Wakeley, J. (2001). Distinguishing migration from isolation: A markov chain monte carlo approach. Genetics, 158(2):885–896.

Nikolic, N. and Chevalet, C. (2014). Detecting past changes of effective population size. Evolutionary Applications, 7(6):663–681.

Nordborg, M. (2001). Coalescent theory. Handbook of Statistical Genetics.

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

Notohara, M. (1993). The strong-migration limit for the genealogical process in geographically structured populations. Journal of Mathematical Biology, 31(2):115–122.

Olivieri, G. L., Sousa, V., Chikhi, L., and Radespiel, U. (2008). From genetic diversity and structure to conservation: genetic signature of recent population declines in three mouse lemur species (microcebus spp.). Biological Conservation, 141(5):1257–1271.

Paz-Vinas, I., Quéméré, E., Chikhi, L., Loot, G., and Blanchet, S. (2013). The demographic history of populations experiencing asymmetric gene flow: combining simulated and empirical data. Molecular ecology, 22(12):3279–3291.

Peter, B. M., Wegmann, D., and Excoffier, L. (2010). Distinguishing between population bottleneck and population subdivision by a bayesian model choice procedure. Molecular Ecology, 19(21):4648–4660.

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.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics, 155(2):945–959.

Prufer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., de Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., Siebauer, M., Theunert, C., Tandon, A., Moorjani, P., Pickrell, J., Mullikin, J. C., Vohr, S. H., Green, R. E., Hellmann, I., Johnson, P. L. F., Blanche, H., Cann, H., Kitzman, J. O., Shendure, J., Eichler, E. E., Lein, E. S., Bakken, T. E., Golovanova, L. V., Doronichev, V. B., Shunkov, M. V., Derevianko, A. P., Viola, B., Slatkin, M., Reich, D., Kelso, J., and Paabo, S. (2014). The complete genome sequence of a neanderthal from the altai mountains. Nature, 505(7481):43–49.

Pybus, O. G., Rambaut, A., and Harvey, P. H. (2000). An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics, 155(3):1429–1437.

Quéméré, E., Amelot, X., Pierson, J., Crouau-Roy, B., and Chikhi, L. (2012). Genetic data suggest a natural prehuman origin of open habitats in northern madagascar and question the deforestation narrative in this region. Proceedings of the National Academy of Sciences, 109(32):13028–13033.

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

Ruegg, A. (1989). Processus stochastiques: avec applications aux phénomènes d'attente et de fiabilité, volume 6. PPUR presses polytechniques.

Salmona, J., Salamolard, M., Fouillot, D., Ghestemme, T., Larose, J., Centon, J.-F., Sousa, V., Dawson, D. A., Thebaud, C., and Chikhi, L. (2012). Signature of a pre-human population decline in the critically endangered reunion island endemic forest bird coracina newtoni. PloS one, 7(8):e43524.

Schiffels, S. and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. Nat Genet, 46(8):919–925.

Sheehan, S., Harris, K., and Song, Y. S. (2013). Estimating variable effective population sizes from multiple genomes: A sequentially markov conditional sampling distribution approach. Genetics, 194(3):647–662.

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.

Sousa, V. C., Beaumont, M. A., Fernandes, P., Coelho, M. M., and Chikhi, L. (2012). Population divergence with or without admixture: selecting models using an abc approach. Heredity, 108(5):521–530.

Städler, T., Haubold, B., Merino, C., Stephan, W., and Pfaffelhuber, P. (2009). The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. Genetics, 182(1):205–216.

Strimmer, K. and Pybus, O. G. (2001). Exploring the demographic history of dna sequences using the generalized skyline plot. Molecular Biology and Evolution, 18(12):2298–2305.

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

Tavaré, S. (2004). Part i: Ancestral inference in population genetics. In Picard, J., editor, Lectures on Probability Theory and Statistics, volume 1837 of Lecture Notes in Mathematics, pages 1–188. Springer Berlin Heidelberg.

Vitti, J. J., Grossman, S. R., and Sabeti, P. C. (2013). Detecting natural selection in genomic data. Annual Review of Genetics, 47:97–120.

Vogl, C., Das, A., Beaumont, M., Mohanty, S., and Stephan, W. (2003). Population subdivision and molecular sequence variation: Theory and analysis of drosophila ananassae data. Genetics, 165(3):1385–1395.

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

Wakeley, J. (2001). The coalescent in an island model of population subdivision with variation among demes. Theoretical Population Biology, 59(2):133 – 144.

Wakeley, J. (2009). Coalescent theory: an introduction. Number 575: 519.2 WAK.

Wakeley, J. and Aliacar, N. (2001). Gene genealogies in a metapopulation. Genetics, 159(2):893–905.

Wakeley, J. and Sargsyan, O. (2009). Extensions of the coalescent effective population size. Genetics, 181(1):341–345.

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., et al. (2008). The diploid genome sequence of an asian individual. Nature, 456(7218):60–65.

Wang, Y. and Hey, J. (2010). Estimating divergence parameters with small samples from a large number of loci. Genetics, 184(2):363–379.

Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. Theoretical Population Biology, 7(2):256 – 276.

Weinberg, W. (1908). Über vererbungsgesetze beim menschen. Zeitschrift für Induktive Abstammungs- und Vererbungslehre, 1(1):440–460.

Wilkins, J. F. and Wakeley, J. (2002). The coalescent in a continuous, finite, linear population. Genetics, 161(2):873–888.

Wilkinson-Herbots, H. M. (1998). Genealogy and subpopulation differentiation under various models of population structure. Journal of Mathematical Biology, 37(6):535–585.

Wright, S. (1931). Evolution in mendelian populations. Genetics, 16(2):97.

Zhan, X., Pan, S., Wang, J., Dixon, A., He, J., Muller, M. G., Ni, P., Hu, L., Liu, Y., Hou, H., et al. (2013). Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle. Nature Genetics, 45(5):563–566.

Zhao, S., Zheng, P., Dong, S., Zhan, X., Wu, Q., Guo, X., Hu, Y., He, W., Zhang, S., Fan, W., et al. (2013). Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. Nature Genetics, 45(1):67–71.

Zhou, X., Sun, F., Xu, S., Fan, G., Zhu, K., Liu, X., Chen, Y., Shi, C., Yang, Y., Huang, Z., Chen, J., Hou, H., Guo, X., Chen, W., Chen, Y., Wang, X., Lv, T., Yang, D., Zhou, J., Huang, B., Wang, Z., Zhao, W., Tian, R., Xiong, Z., Xu, J., Liang, X., Chen, B., Liu, W., Wang, J., Pan, S., Fang, X., Li, M., Wei, F., Xu, X., Zhou, K., Wang, J., and Yang, G. (2013). Baiji genomes reveal low genetic variability and new insights into secondary aquatic adaptations. Nat Commun, 4:–.

Zhou, X., Wang, B., Pan, Q., Zhang, J., Kumar, S., Sun, X., Liu, Z., Pan, H., Lin, Y., Liu, G., et al. (2014). Whole-genome sequencing of the snub-nosed monkey provides insights into folivory and evolutionary history. Nature Genetics, 46(12):1303–1310.

RÉSUMÉ : Le développement des nouvelles techniques de séquençage élargit l' horizon de la génétique de populations. Une analyse appropriée des données génétiques peut augmenter notre capacité à reconstruire l'histoire des populations. Cette énorme quantité de données disponibles peut aider les chercheurs en biologie et anthropologie à mieux estimer les changements démographiques subis par une population au cours du temps, mais induit aussi de nouveaux défis. Lorsque les modèles sous-jacents sont trop simplistes il existe un risque très fort d'être amené à des conclusions erronées sur la population étudiée. Il a été montré que certaines caractéristiques présentes dans l'ADN des individus d'une population structurée se trouvent aussi dans l'ADN de ceux qui proviennent d'une population sans structure dont la taille a changé au cours du temps. Par conséquent il peut s'avérer très difficile de déterminer si les changements de taille inférés à partir des données génétiques ont vraiment eu lieu ou s'il s'agit simplement des effets liés à la structure. D'ailleurs la quasi totalité des méthodes pour inférer les changements de taille d'une population au cours du temps sont basées sur des modèles qui négligent la structure.
Dans cette thèse, de nouveaux résultats de génétique de populations sont présentés. Premièrement, nous présentons une méthodologie permettant de faire de la sélection de modèle à partir de l'ADN d'un seul individu diploïde. Cette première étude se limite à un modèle simple de population non structurée avec un changement de taille et à un modèle considérant une population de taille constante mais structurée. Cette nouvelle méthode utilise la distribution des temps de coalescence de deux gènes pour identifier le modèle le plus probable et ouvre ainsi la voie pour de nouvelles méthodes de sélection de modèles structurés et non structurés, à partir de données génomiques issues d'un seul individu. Deuxièmement, nous montrons, par une ré-interprétation du taux de coalescence que, pour n'importe quel scénario structuré, et plus généralement n'importe quel modèle, il existe toujours un scénario considérant une population panmictique avec une fonction précise de changements de taille dont la distribution des temps de coalescence de deux gènes est identique a celle du scénario structuré. Cela non seulement explique pourquoi les méthodes d'inférence démographique détectent souvent des changements de taille n'ayant peut-être jamais eu lieu, mais permet aussi de prédire les changements de taille qui seront reconstruits lorsque des méthodes basées sur l'hypothèse de panmixie sont appliquées à des données issues de scénarios plus complexes. Finalement, une nouvelle approche basée sur un processus de Markov est développée et permet de caractériser la distribution du temps de coalescence de deux gènes dans une population structurée soumise à des événements démographiques tel que changement de flux de gènes et changements de taille. Une discussion est menée afin de décrire comment cette méthode donne la possibilité de reconstruire l'histoire démographique à partir de données génomiques tout en considérant la structure.

MOTS-CLEFS : génétique des populations, théorie de la coalescence, temps de coalescence, histoire démographique, chaîne de Markov, estimation par maximum de vraisemblance.

ABSTRACT: The rapid development of DNA sequencing technologies is expanding the horizons of population genetic studies. It is expected that genomic data will increase our ability to reconstruct the history of populations. While this increase in genetic information will likely help biologists and anthropologists to reconstruct the demographic history of populations, it also poses big challenges. In some cases, simplicity of the model may lead to erroneous conclusions about the population under study. Recent works have shown that DNA patterns expected in individuals coming from structured populations correspond with those of unstructured populations with changes in size through time. As a consequence it is often difficult to determine whether demographic events such as expansions or contractions (bottlenecks) inferred from genetic data are real or due to the fact that populations are structured in nature. Moreover, almost no inferential method allowing to reconstruct past demographic size changes takes into account structure effects.
In this thesis, some recent results in population genetics are presented: (i) a model choice procedure is proposed to distinguish one simple scenario of population size change from one of structured population, based on the coalescence times of two genes, showing that for these simple cases, it is possible to distinguish both models using genetic information form one single individual; (ii) by using the notion of instantaneous coalescent rate, it is demonstrated that for any scenario of structured population or any other one, regardless how complex it could be, there always exists a panmitic scenario with a precise function of population size changes having exactly the same distribution for the coalescence times of two genes. This not only explains why spurious signals of bottlenecks can be found in structured populations but also predicts the demographic history that actual inference methods are likely to reconstruct when applied to non panmitic populations. Finally, (iii) a method based on a Markov process is developed for inferring past demographic events taking the structure into account. This is method uses the distribution of coalescence times of two genes to detect past demographic changes in structured populations from the DNA of one single individual. Some applications of the model to genomic data are discussed.

KEY-WORDS: population genetics, coalescence theory, coalescence time, demographic history, Markov chain, maximum likelihood estimation.