Theoretical Population Biology 104 (2015) 46–58

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

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

Olivier Mazet [a], Willy Rodríguez [a], Lounès Chikhi [b,c,d,*]

[a] UMR 5219, Institut de Mathématiques de Toulouse, Université de Toulouse & CNRS, France
[b] CNRS, Université Paul Sabatier, ENFA, UMR 5174 EDB (Laboratoire Évolution & Diversité Biologique), F-31062 Toulouse, France
[c] Université de Toulouse, UPS, EDB, F-31062 Toulouse, France
[d] Instituto Gulbenkian de Ciência, P-2780-156 Oeiras, Portugal

ARTICLE INFO

Article history:
Received 30 January 2015
Available online 25 June 2015

Keywords:
Symmetric island model
Population size change
Maximum likelihood estimation
Demographic history
Coalescence time

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 analyze 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, and (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.

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 increasing 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

* Correspondence to: CNRS, Université Paul Sabatier, Laboratoire Evolution & Diversité Biologique, Bâtiment 4R1, 118 route de Narbonne, 31062 Toulouse cedex 9, France.
E-mail address: lounes.chikhi@univ-tlse3.fr (L. Chikhi).

http://dx.doi.org/10.1016/j.tpb.2015.06.003
0040-5809/

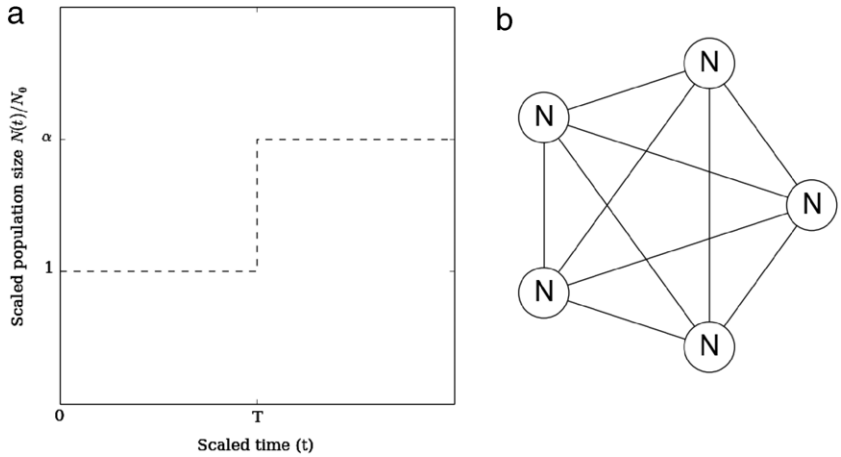

Fig. 1. Demographic models. (a) Single step population size change (SSPSC) model. The x-axis represents t, the time to the past in units of generations scaled by the number of genes. At time $t = T$, (going from the present to the past) the population size changes instantaneously from N0 to N1 by a factor 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 $N_0 = N$.

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., 2006a; 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 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. Methods

2.1. Demographic models

2.1.1. 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 (Fig. 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 (Fig. 1) whereas if $\alpha < 1$ it expanded. Since

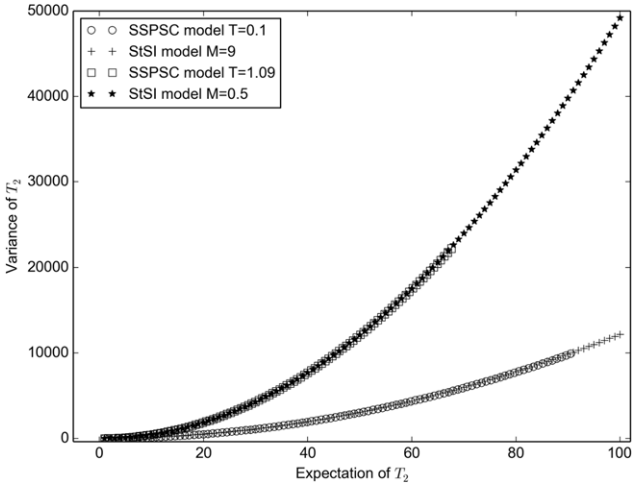

Fig. 2. Expected value and Variance of T_2 under the SSPSC and StSI models. This figure illustrates how both models can have the same pair of values $(E(T_2), Var(T_2))$ for many sets of parameters. For the SSPSC model the time at which the population size change occurred was fixed to $T = 0.1$ whereas 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.

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.

2.1.2. Structured population

Here we consider the classical symmetric *n-island* model Wright (1931), see Fig. 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.

2.2. 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 (Figs. 2 and 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.

2.2.1. 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

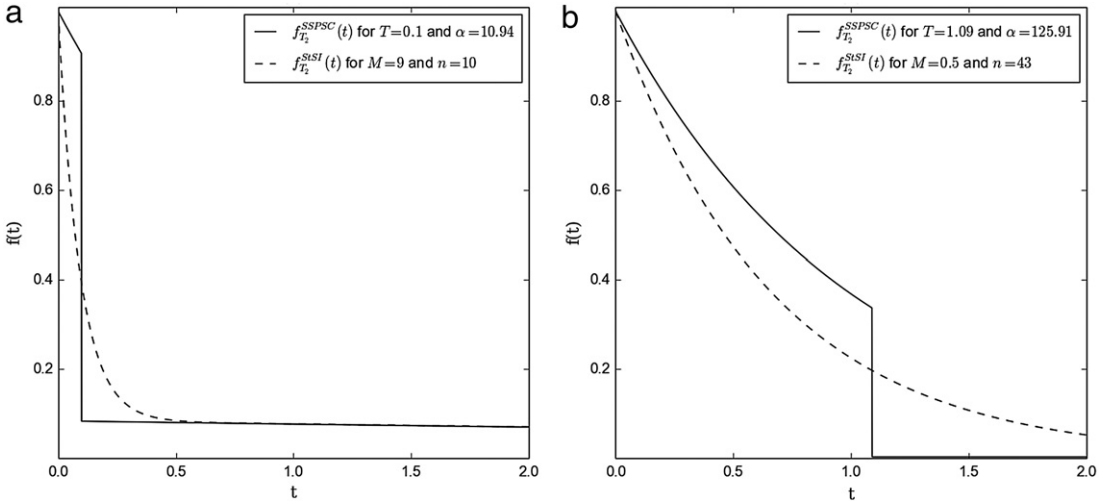

Fig. 3. Density of T_2 under the SSPSC and StSI models. Two sets of parameter values (panels (a) and (b), respectively) were chosen on the basis that expectations and variances were close. Panel (a) Density for the SSPSC model with $T = 0.1$ and $\alpha = 10.94$, and for the StSI model with $M = 9$ and $n = 10$. 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 $Var\left(T_2^{StSI}\right) = 8905$.

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.

2.2.2. Derivation of the distribution of coalescence times

The distribution of T_2^{SSPSC}. The generalization 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 (see Appendix A). In the case of the *SSPSC* model, this leads to the following *pdf*:

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

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 (Appendix A) 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. First moments

Eqs. (1) and (2) 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:

$$\mathbb{E}\left(T_2^{SSPSC}\right) = 1 + e^{-T}(\alpha - 1),$$

$$Var\left(T_2^{SSPSC}\right) = 1 + 2Te^{-T}(\alpha-1) + 2\alpha e^{-T}(\alpha-1) - (\alpha-1)^2 e^{-2T},$$

$$\mathbb{E}(T_2^{StSI}) = n,$$

$$Var(T_2^{StSI}) = n^2 + \frac{2(n-1)^2}{M}.$$

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 Fig. 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 behaviors. 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 illustrate 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, Fig. 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 paper when we want to mention the number of independent T_2 values.

2.3. Model choice and parameter estimation

2.3.1. General principle and parameter combinations

Given a sample $(t_1, \ldots, t_{n_L})$ 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 an 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, Appendix A 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 $(t_1, \ldots, t_{n_L})$ 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 $(t_1, \ldots, t_{2n_L})$ 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, 20 000). 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.

2.3.2. Maximum likelihood estimation (MLE) in the SSPSC case

We know from Eq. (1) 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 $t = (t_1, t_2, \ldots, t_{n_L})$, the likelihood is:

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

and taking the *log* it gives:

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

Lemma 2.1. *Given a set of n_L independent observations $\{t_1, t_2, \ldots, t_{n_L}\}$, 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*

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

with

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

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 $\{(\alpha, t_a),\ \alpha \in \mathbb{R}^+\},\ a \in \{1, 2, \ldots, n_L\}$. 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: $(\hat{\alpha}, \hat{T}) = \text{argmax}_{a\in\{1,\ldots,n_L\}}\{\log(\mathbb{L}_{SSPSC}(m_a))\}$.
For the proof and some comments, see Supplementary materials (Appendix A).

2.3.3. MLE in the StSI case

Under the StSI model the expression of the critical points is not analytically derived. We know from Eq. (2) the *pdf* of coalescence times for two genes. Given n_L independent values $t = (t_1, t_2, \ldots, t_{n_L})$ 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 of them had the largest log-likelihood value.

2.3.4. Akaike Information Criterion and robustness to model departures

Once we have computed our approximations 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, Appendix A) 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

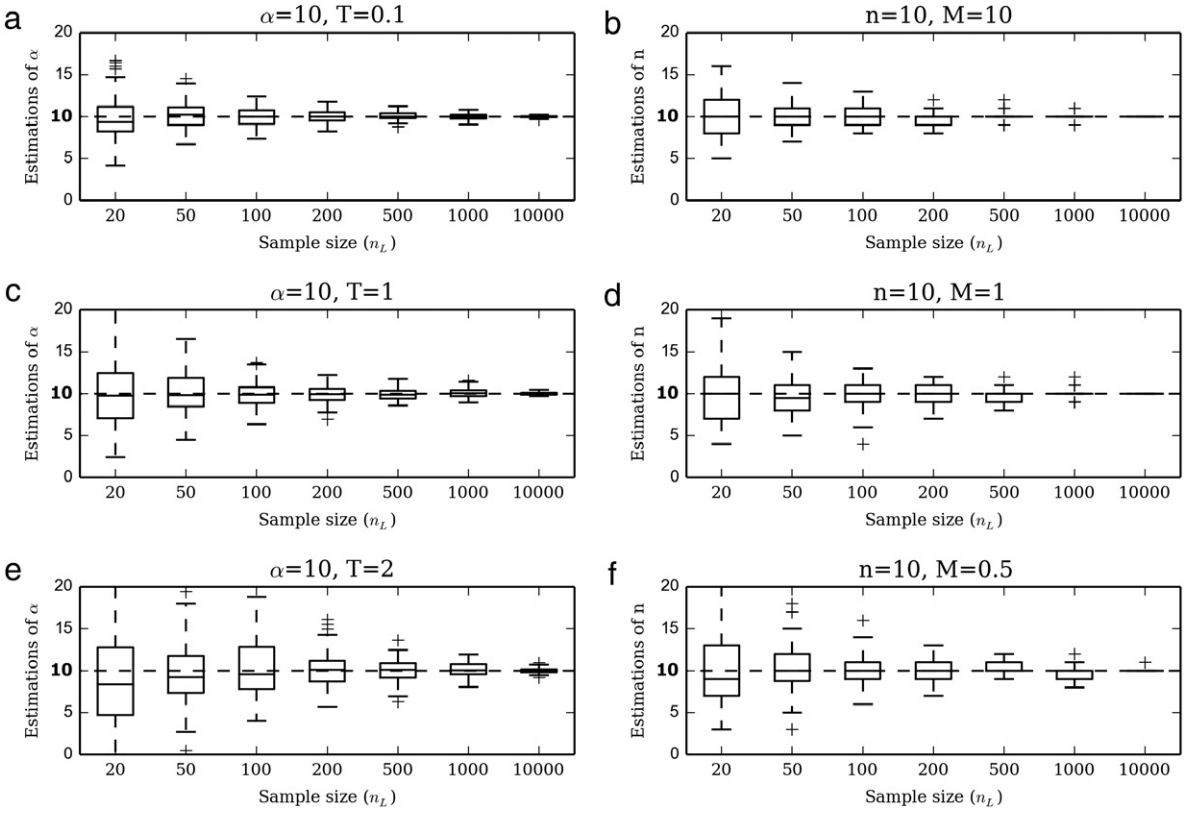

Fig. 4. Estimation of α and n. Panels (a), (c) and (e) Estimation of α under the SSPSC model for different sample sizes and T values. Simulations performed with $\alpha = 10$ and $T = (0.1, 1, 2)$. Panels (b), (d) and (f) Estimation of n under the StSI model for different sample sizes and M values. Simulations performed with $n = 10$ and $M = (10, 1, 0.5)$.

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).

3. Results

Fig. 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 materials (Appendix A) for 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 materials (Appendix A) for 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, Appendix A 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.

Fig. 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 materials (Appendix A) for 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 materials (Appendix A) for 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 Fig. 6 we show some results of the *KS* test for the two cases (see the Supplementary materials, Appendix A for other parameter combinations). In the left-hand panels ((a), (c), and (e)) the data

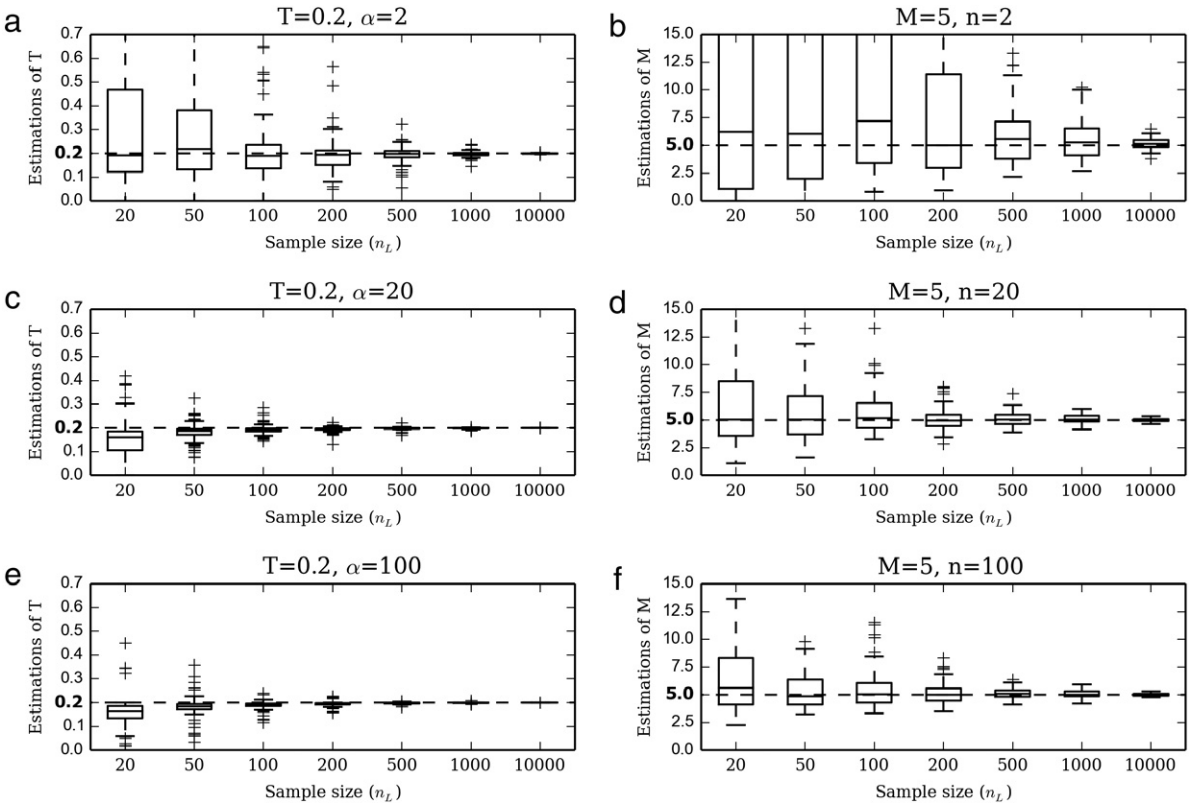

Fig. 5. Estimation of T and M. Panels (a), (c), (e) Estimation of T under the SSPSC model for different sample sizes and values of α. Simulations performed with $\alpha = (2, 20, 100)$ and $T = 0.2$. Panels (b), (d), (f) Estimation of M under the StSI model for different sample sizes and values of n.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

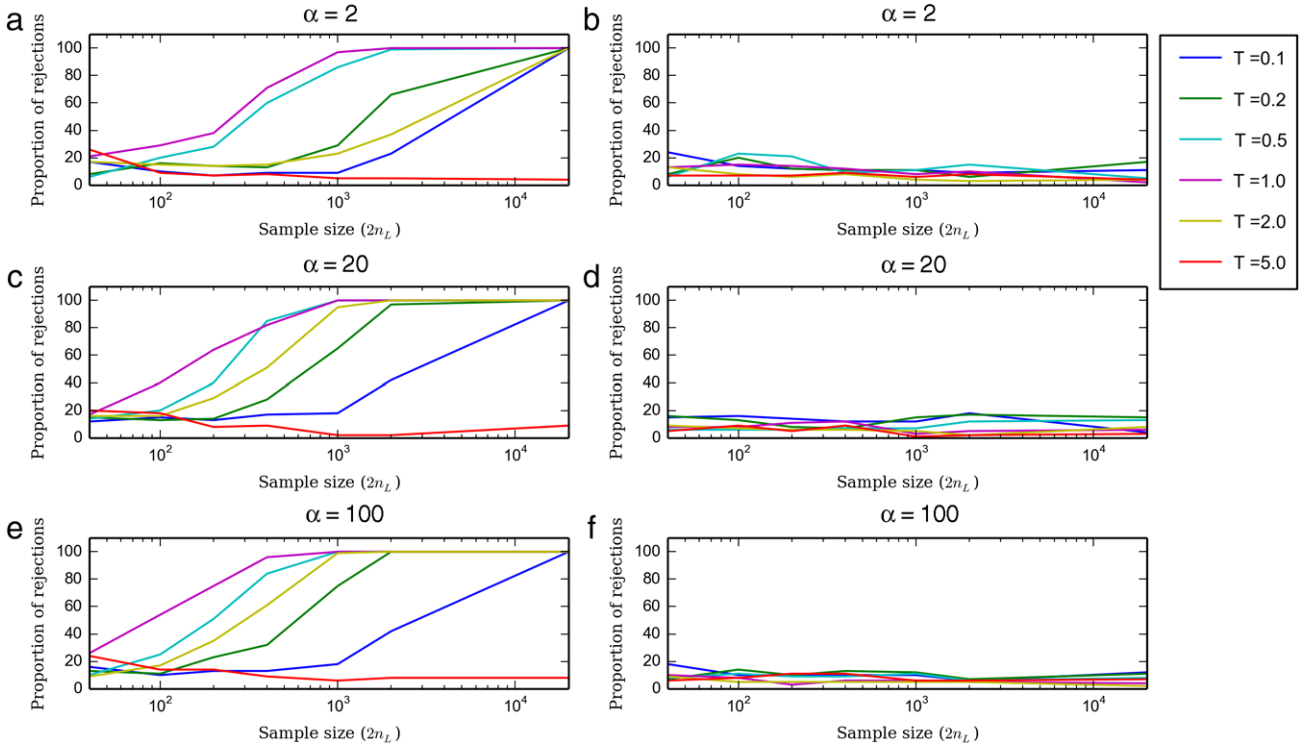

Fig. 6. Proportion of rejected data sets simulated under the SSPSC model. Panels (a), (c) and (e) the reference model is the StSI model. Panels (b), (d), and (f) the reference model is the SSPSC, *i.e.* the model under which the data were simulated. Note that for the abscissa we used $2n_L$ instead of 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.

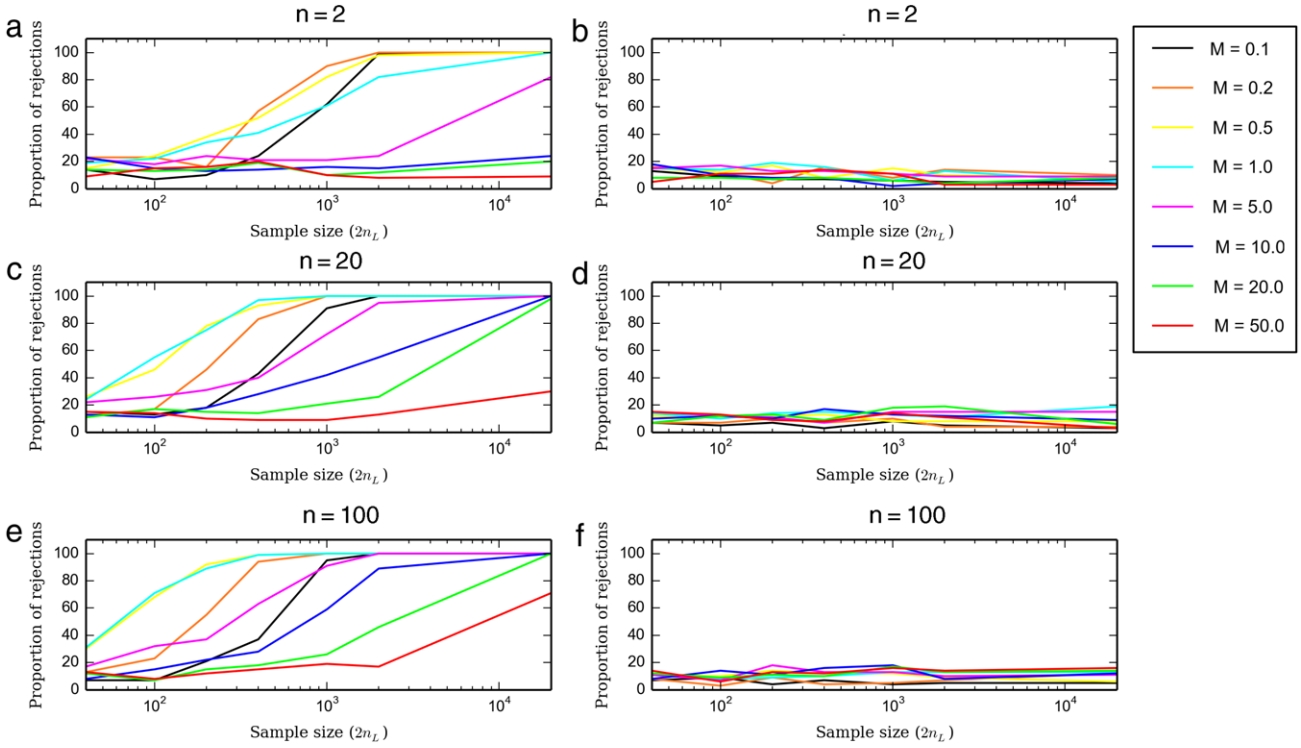

Fig. 7. Proportion of rejected data sets simulated under the StSI model. Panels (a), (c), and (e) the reference model is the SSPSC. Panels (b), (d), and (f) the reference model is the StSI model, *i.e.* the model under which the data were simulated. Note that for the abscissa we used $2n_L$ instead of 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.

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 Figs. 4 and 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) be. 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 10,000, only *ca.* 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 is. 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, Appendix A), 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 an 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.

Fig. 7 is similar to Fig. 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 Fig. 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 (Figs. 6 and 7) show that it is possible to identify the model under which the data were generated by using a single diploid individual.

Fig. 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"

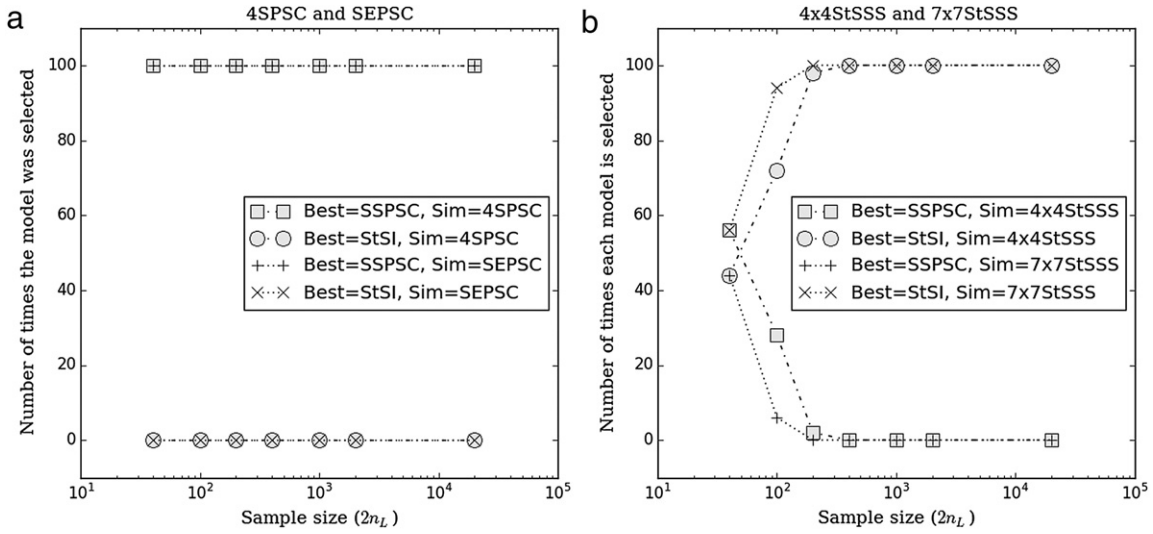

Fig. 8. Model choice using the AIC for various demographic models. This figure shows the proportion of times the AIC selected the SSPSC (resp. the StSI) as the best model, as a function of n_L, the number of independent loci or T_2 values. For each n_L value, the experiment was repeated 100 times, and 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, the 4x4StSSS (with 4×4 islands) and the 7x7StSSS (with 7×7 islands). This figure shows that the AIC provides very good results to identify a structured model compared to a model of population size change.

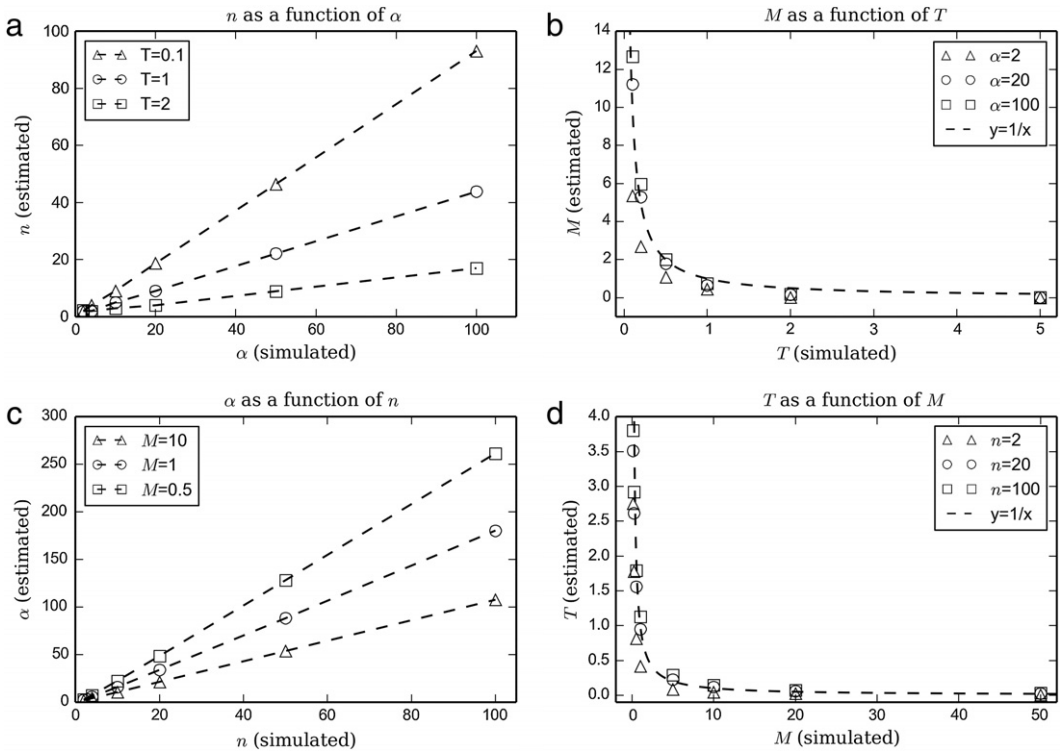

Fig. 9. Relationships between parameters of the models.

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, Appendix A). 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.

Fig. 9 is divided into 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 in 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.

4. Discussion

In this study we have analyzed the distribution of coalescence times under two simple demographic models. We have shown that even though these demographic models are strikingly different (Fig. 1) there is always a way to find parameter values for which both models will have the same first two moments (Fig. 2). We have also shown that there are intrinsic relationships between the parameters of the two models (Fig. 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 an *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 demographic 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 effective 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.

4.1. 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, Appendix A) 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

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

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.

4.2. 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, Appendix A 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 measured 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 1000 and 10,000 will be enough to estimate the models' parameters and to identify or reject models with great confidence.

4.3. 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 neighboring 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.

4.4. 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 recognized in the last decade the importance of population structure as potential factor biasing inference of demographic (Leblois et al., 2006b; 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.

4.5. 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 asymmetrical gene flow between populations.

4.6. 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 the best of 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).

Acknowledgments

We are grateful to S. Boitard and S. Grusea for numerous and fruitful discussions and inputs throughout the development of this work. We thank I. Paz, I. Pais, J. Salmona, J. Chave for discussion and helpful comments that improved the clarity of the text. We also thank B. Laurent for her comments on the KS test and for various discussions on the results and statistical issues. We thank J. Howard and the Fundação Calouste Gulbenkian for their continuous support. We are grateful to the genotoul bioinformatics platform Toulouse Midi-Pyrenees for providing computing and storage resources. This work was partly performed using HPC resources from CALMIP (Grant 2012—projects 43 and 44) from Toulouse, France. This study was funded by the Fundação para a Ciência e Tecnologia (Ref. PTDC/BIA-BIC/4476/2012), the Projets Exploratoires Pluridisciplinaires (PEPS 2012 Bio-Maths-Info) project, the LABEX entitled TULIP (ANR-10-LABX-41) as well as the Pôle de Recherche et d'Enseignement Supérieur (PRES) and the Région Midi-Pyrénées, France. We are also grateful for the critical and helpful comments of the (not so) anonymous reviewers and to the Genetics and TPB editors (J. Wall and N. Rosenberg) who facilitated the submission process and helped us to improve the clarity and quality of the paper.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.tpb.2015.06.003.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control 19 (6), 716–723.

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. Annu. Rev. Ecol. Evol. Syst. 41, 379–406.

Beaumont, M.A., Zhang, W., Balding, D.J., 2002. Approximate Bayesian computation in population genetics. Genetics 162 (4), 2025–2035.

Broquet, T., Angelone, S., Jaquiery, J., Joly, P., Lena, J.-P., Lengagne, T., Plenet, S., Luquet, E., Perrin, N., 2010. Genetic bottlenecks driven by population disconnection. Conserv. Biol. 24 (6), 1596–1605.

Chen, C., Durand, E., Forbes, F., François, O., 2007. Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. Mol. Ecol. Notes 7 (5), 747–756.

Chikhi, L., Sousa, V.C., Luisi, P., Goossens, B., 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., 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., Estoup, A., 2008. Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. Bioinformatics 24 (23), 2713–2719.

Currat, M., Excoffier, L., Maddison, W., Otto, S.P., Ray, N., Whitlock, M.C., 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), 172.

Donnelly, P., Tavaré, S., 1995. Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. 29 (1), 401–421.

Durrett, R., 2008. Probability Models for DNA Sequence Evolution. Springer.

Girod, C., Vitalis, R., Leblois, R., Fréville, H., 2011. Inferring population decline and expansion from microsatellite data: A simulation-based evaluation of the msvar method. Genetics 188 (1), 165–179. URL: http://www.genetics.org/content/188/1/165.abstract.

Goossens, B., Chikhi, L., Ancrenaz, M., Lackman-Ancrenaz, I., Andau, P., Bruford, M.W., 2006. Genetic signature of anthropogenic population collapse in orang-utans. PLoS Biol. 4 (2), e25.

Griffiths, R., 1981. The number of heterozygous loci between two randomly chosen completely linked sequences of loci in two subdivided population models. J. Math. Biol. 12 (2), 251–261.

Guillot, G., Mortier, F., Estoup, A., 2005. GENELAND: a computer package for landscape genetics. Mol. Ecol. Notes 5 (3), 712–715.

Hallatschek, O., Fisher, D.S., 2014. Acceleration of evolutionary spread by long-range dispersal. Proc. Natl. Acad. Sci. 111 (46), E4911–E4919.

Heller, R., Chikhi, L., 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 (Ph.D. thesis),
Hudson, R.R., 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18 (2), 337–338. URL: http://bioinformatics.oxfordjournals.org/content/18/2/337.abstract.
Hudson, R.R., et al., 1990. Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. 7 (1), 44.
Jones, E., Oliphant, T., Peterson, P., et al. 2001. SciPy: Open source scientific tools for Python. http://www.scipy.org/ [Online; accessed 18.11.14].
Kimura, M., Weiss, G.H., 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics 49 (4), 561–576.
Leblois, R., Estoup, A., Streiff, R., 2006a. Genetics of recent habitat contraction and reduction in population size: does isolation by distance matter? Mol. Ecol. 15 (12), 3601–3615.
Leblois, R., Estoup, A., Streiff, R., 2006b. Genetics of recent habitat contraction and reduction in population size: does isolation by distance matter? Mol. Ecol. 15 (12), 3601–3615.
Li, H., Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. Nature 475 (7357), 493–496.
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., Hammer, M., 2015. Inference of Gorilla demographic and selective history from whole genome sequence data. Mol. Biol. Evol. 600–612.
Nei, M., Takahata, N., 1993. Effective population size, genetic diversity, and coalescence time in subdivided populations. J. Mol. Evol. 37 (3), 240–244.
Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. Comput. J. 7 (4), 308–313.
Nielsen, R., Beaumont, M.A., 2009. Statistical inferences in phylogeography. Mol. Ecol. 18 (6), 1034–1047.
Olivieri, G.L., Sousa, V., Chikhi, L., Radespiel, U., 2008. From genetic diversity and structure to conservation: genetic signature of recent population declines in three mouse lemur species (Microcebus spp.). Biol. Conserv. 141 (5), 1257–1271.
Paz-Vinas, I., Quéméré, E., Chikhi, L., Loot, G., Blanchet, S., 2013. The demographic history of populations experiencing asymmetric gene flow: combining simulated and empirical data. Mol. Ecol. 22 (12), 3279–3291.
Peter, B.M., Wegmann, D., Excoffier, L., 2010. Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. Mol. Ecol. 19 (21), 4648–4660.
Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155 (2), 945–959.
Quéméré, E., Amelot, X., Pierson, J., Crouau-Roy, B., Chikhi, L., 2012. Genetic data suggest a natural prehuman origin of open habitats in northern Madagascar and question the deforestation narrative in this region. Proc. Natl. Acad. Sci. 109 (32), 13028–13033.
Rogers, A.R., Harpending, H., 1992. Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. 9 (3), 552–569.
Salmona, J., Salamolard, M., Fouillot, D., Ghestemme, T., Larose, J., Centon, J.-F., Sousa, V., Dawson, D.A., Thebaud, C., 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., Durbin, R., 2014. Inferring human population size and separation history from multiple genome sequences. Nature Genet..
Sousa, V.C., Beaumont, M.A., Fernandes, P., Coelho, M.M., Chikhi, L., 2012. Population divergence with or without admixture: selecting models using an ABC approach. Heredity 108 (5), 521–530. URL: http://dx.doi.org/10.1038/hdy.2011.116.
Städler, T., Haubold, B., Merino, C., Stephan, W., Pfaffelhuber, P., 2009. The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. Genetics 182 (1), 205–216.
Tavaré, S., 2004. Part I: Ancestral inference in population genetics. In: Lectures on Probability Theory and Statistics. Springer, pp. 1–188.
Vitti, J.J., Grossman, S.R., Sabeti, P.C., 2013. Detecting natural selection in genomic data. Annu. Rev. Genet. 47, 97–120.
Wakeley, J., 1999. Nonequilibrium migration in human history. Genetics 153 (4), 1863–1871.
Wright, S., 1931. Evolution in Mendelian populations. Genetics 16 (2), 97.