



HAL
open science

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

Olivier Mazet, Willy Rodríguez, Lounès Chikhi

► **To cite this version:**

Olivier Mazet, Willy Rodríguez, Lounès Chikhi. Demographic inference using genetic data from a single individual: separating population size variation from population structure. *Theoretical Population Biology*, 2015, 104, pp.46 - 58. 10.1016/j.tpb.2015.06.003 . hal-01087849

HAL Id: hal-01087849

<https://hal.science/hal-01087849>

Submitted on 27 Nov 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Open licence - etalab

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

Mazet Olivier^{*}, Rodríguez Valcarce Willy^{*}, and Chikhi Lounès^{§,†,‡}

^{*}Université de Toulouse, Institut National des Sciences Appliquées, Institut de
Mathématiques de Toulouse, F-31077 Toulouse, France

[§]CNRS, Université Paul Sabatier, ENFA, UMR 5174 EDB (Laboratoire Évolution &
Diversité Biologique), F-31062 Toulouse, France

[†]Université de Toulouse, UPS, EDB, F-31062 Toulouse, France

[‡]Instituto Gulbenkian de Ciência, P-2780-156 Oeiras, Portugal

Running Head: Structure vs Population Size Change

Key Words: symmetric island model, population size change, maximum likelihood estimation, demographic history, coalescence time

Corresponding Author:

Chikhi Lounès

CNRS, Université Paul Sabatier, Laboratoire Evolution & Diversité Biologique, Bâtiment
4R1, 118 route de Narbonne, 31062 Toulouse cedex 9, France.

`lounes.chikhi@univ-tlse3.fr`

Abstract

1
2 The rapid development of sequencing technologies represents new opportunities for pop-
3 ulation genetics research. It is expected that genomic data will increase our ability to re-
4 construct the history of populations. While this increase in genetic information will likely
5 help biologists and anthropologists to reconstruct the demographic history of populations,
6 it also represents new challenges. Recent work has shown that structured populations gen-
7 erate signals of population size change. As a consequence it is often difficult to determine
8 whether demographic events such as expansions or contractions (bottlenecks) inferred from
9 genetic data are real or due to the fact that populations are structured in nature. Given
10 that few inferential methods allow us to account for that structure, and that genomic data
11 will necessarily increase the precision of parameter estimates, it is important to develop new
12 approaches. In the present study we analyse two demographic models. The first is a model
13 of instantaneous population size change whereas the second is the classical symmetric island
14 model. We (i) re-derive the distribution of coalescence times under the two models for a sam-
15 ple of size two, (ii) use a maximum likelihood approach to estimate the parameters of these
16 models (iii) validate this estimation procedure under a wide array of parameter combina-
17 tions, (iv) implement and validate a model choice procedure by using a Kolmogorov-Smirnov
18 test. Altogether we show that it is possible to estimate parameters under several models and
19 perform efficient model choice using genetic data from a single diploid individual.

INTRODUCTION

20

21 The sheer amount of genomic data that is becoming available for many organisms with the
22 rapid development of sequencing technologies represents new opportunities for population
23 genetics research. It is hoped that genomic data will increase our ability to reconstruct the
24 history of populations (LI and DURBIN 2011) and detect, identify and quantify selection
25 (VITTI *et al.* 2013). While this increase in genetic information will likely help biologists
26 and anthropologists to reconstruct the demographic history of populations, it also exposes
27 old challenges in the field of population genetics. In particular, it becomes increasingly
28 necessary to understand how genetic data observed in present-day populations are influenced
29 by a variety of factors such as population size changes, population structure and gene flow
30 (NIELSEN and BEAUMONT 2009). Indeed, the use of genomic data does not necessary
31 lead to an improvement of statistical inference. If the model assumed to make statistical
32 inference is fundamentally mis-specified, then increasing the amount of data will lead to
33 increased precision for perhaps misleading if not meaningless parameters and will not reveal
34 new insights (NIELSEN and BEAUMONT 2009; CHIKHI *et al.* 2010; HELLER *et al.* 2013).

35 For instance, several recent studies have shown that the genealogy of genes sampled from
36 a deme in an island model is similar to that of genes sampled from a non structured isolated
37 population submitted to a demographic bottleneck (PETER *et al.* 2010; CHIKHI *et al.* 2010;
38 HELLER *et al.* 2013). As a consequence, using a model of population size change for a
39 spatially structured population may falsely lead to the inference of major population size
40 changes (STÄDLER *et al.* 2009; PETER *et al.* 2010; CHIKHI *et al.* 2010; HELLER *et al.* 2013;
41 PAZ-VINAS *et al.* 2013). Conversely, assuming a structured model to estimate rates of gene
42 flow when a population has been submitted to a population size change, may also generate
43 misleading conclusions, even though the latter case has been much less documented. More
44 generally, previous studies have shown that spatial processes can mimic selection (CURRAT
45 *et al.* 2006), population size changes (CHIKHI *et al.* 2010; HELLER *et al.* 2013) or that changes
46 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 (PRITCHARD *et al.* 2000; 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 individual. 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 it may still be important to improve our understanding of the coalescent under structured models.

In the present study we are interested in describing the properties of the coalescent under two models of population size change and population structure, respectively, and in devising a new statistical test and estimation procedures. More specifically we re-derive 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 use a maximum likelihood approach to estimate the parameters of interest for each model (timing and ratio of population size change former and rate of gene flow and deme size for the latter). We develop a statistical test that identifies data sets generated under the two models. Finally, we discuss how these results may apply to genomic data and how they could be extended to real data sets (since T_2 is not usually known) and 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

⁷⁴ extensively used on genomic data and also uses a sample size of two.

76 **Demographic models**

77 *Population size change:* We consider a simple model of population size change, where $N(t)$
 78 represents the population size (N , in units of genes or haploid genomes) as a function of
 79 time (t) expressed in generations scaled by N , the population size, and where $t = 0$ is the
 80 present, and positive values represent the past (Figure 1 (a)). More specifically we assume a
 81 sudden change in population size at time T in the past, where N changes instantaneously by
 82 a factor α . This can be summarized as $N(t) = N(0) = N_0$ for $t \in [0, T[$, $N(t) = N(T) = \alpha N_0$
 83 for $t \in [T, +\infty[$. If $\alpha > 1$ the population went through a bottleneck (Figure 1) whereas if
 84 $\alpha < 1$ it expanded. Since N represents the population size in terms of haploid genomes,
 85 the number of individuals will therefore be $N/2$ for diploid species. Note also that for a
 86 population of constant size the expected coalescence time of two genes is N generations,
 87 which therefore corresponds to $t = 1$. We call this model the SSPSC, which stands for Single
 88 Step Population Size Change.

89 [Figure 1 about here.]

90 *Structured population:* Here we consider the classical symmetric *n-island* model (WRIGHT
 91 1931), see Figure 1 (b), where we have a set of n islands (or demes) of constant size N ,
 92 interconnected by gene flow with a migration rate m , where $M = Nm$ is the number of im-
 93 migrants (genes) in each island every generation. The whole metapopulation size is therefore
 94 nN (this is the total number of genes, not the effective size). Again, N is the number of
 95 haploid genomes, and $N/2$ the number of diploid individuals. We call this model the StSI,
 96 which stands for Structured Symmetrical Island model.

97 **The distribution of coalescence times: qualitative and quantitative analyses**

98 In this section we used previous results (HERBOTS 1994; DONNELLY and TAVARE 1995)
99 to derive the distribution of coalescent times for the two models of interest. We show
100 that even though they are different, these distributions can be similar under an indefinitely
101 large number of parameter values (Figures 2 and 3). Moreover we show that even when the
102 distributions are distinguishable, their first moments may not be. In particular, we show that
103 the first two moments (mean and variance) are near identical for a large number of parameter
104 combinations. Before doing that we start by providing a simple intuitive rationale explaining
105 why and how a model of population structure can be mistaken for a model of population
106 size change. This intuitive approach is important because it allows us to understand how
107 the parameters of the two models ((T, α) and (M, n) , respectively) are linked.

108 *Intuitive and qualitative rationale:* We start by taking two genes sampled in the present-day
109 population under the Single Step Population Size Change (SSPSC) model. If we assume that
110 $\alpha > 1$ (population bottleneck from an ancient population of size N_1 to a current population
111 of size N_0 , with $N_1 = \alpha N_0$) the probability that the two genes coalesce will vary with time
112 as a function of N_0 , N_1 and T . If T is very small, then most genes will coalesce at a rate
113 determined by N_1 , whereas if T is very large the coalescence rate will be mostly determined
114 by N_0 . If we now take two genes sampled from the same island in the Structured Symmetrical
115 Island (StSI) model, we can also see that their coalescence rate will depend on N , the size
116 of the island and on m , the migration rate. If m is very low, the coalescence rate should
117 mostly depend on N . If m is high, the two genes may see their lineages in different islands
118 before they coalesce. As a consequence the coalescence rate will depend on the whole set of
119 islands and therefore on the product nN , where n is the total number of islands.

120 This intuitive description suggests that there is an intrinsic relationship between T and
121 $1/M$, and between α and n . The reason why structured populations exhibit signals of

122 bottlenecks is because in the recent past the coalescence rate depends on the local island size
 123 N , whereas in a more distant past it depends on nN . In other words, it is as if the population
 124 size had been reduced by a factor of n . As we will see this rationale is only qualitatively
 125 correct, but it suggests that if we want to distinguish them it may be necessary to derive
 126 the full distribution of the coalescence times under the two models. We shall denote these
 127 coalescence times T_2^{SSPSC} and T_2^{StSI} , respectively.

128 *Derivation of the distribution of coalescence times:*

129 **The distribution of T_2^{SSPSC} :** The generalisation of the coalescent in populations of vari-
 130 able size was first rigorously treated in DONNELLY and TAVARE (1995)), and is clearly
 131 exposed in TAVARÉ (2004)). If we denote by $\lambda(t)$ the ratio $\frac{N(t)}{N(0)}$ where t is the time scaled by
 132 the number of genes (*i.e.* units of coalescence time, corresponding to $\lfloor N(0)t \rfloor$ generations),
 133 we can compute the probability density function (*pdf*) $f_{T_2}^{SSPSC}(t)$ of the coalescence time
 134 T_2^{SSPSC} of two genes sampled in the present-day population. Indeed, the probability that
 135 two genes will coalesce at a time greater than t is

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

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

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

136 This leads to the following *pdf*

$$f_{T_2}^{SSPSC}(t) = e^{-t}\mathbb{I}_{[0,T[}(t) + \frac{1}{\alpha}e^{-T-\frac{1}{\alpha}(t-T)}\mathbb{I}_{[T,+\infty[}(t). \quad (2)$$

137 **The distribution of T_2^{StSI} :** Following HERBOTS (1994)), an easy way to derive the dis-
 138 tribution of the coalescence time T_2^{StSI} of two genes for our structured model, is to compute
 139 the probability that two genes are identical by descent when they are sampled from the same
 140 or from different populations. These two probabilities are respectively denoted by $p_s(\theta)$ and
 141 $p_d(\theta)$, where $\theta = 2uN$ is the scaled mutation rate, u being the *per* locus mutation rate.

Indeed, using a classical scaling argument (see for instance TAVARÉ (2004), page 34), we can note that

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

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

143 We can compute this probability as follows. Taking two genes from the same island and
 144 going back in time, there are three events that may occur: a coalescence event (with rate 1),
 145 a mutation event (with rate θ) and a migration event (with rate M). Taking now two genes
 146 from different islands, they cannot coalesce and therefore only a mutation or a migration
 147 event may occur. Migration events can then bring the lineages in the same island with
 148 probability $\frac{1}{n-1}$, and in different islands with probability $\frac{n-2}{n-1}$. We thus obtain the following
 149 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

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

150

where

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

and

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

Noting now that for any θ and any α we have

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

151

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

$$f_{T_2}^{StSI}(t) = ae^{-\alpha t} + (1-a)e^{-\beta t}. \quad (3)$$

152

First moments: Equations 2 and 3 are different hence showing that it is in principle possible

153

to identify genetic data produced under the two demographic models of interest. The two

154

equations can be used to derive the expectation and variance of the two random variables

155

of interest, T_2^{SSPSC} and T_2^{StSI} . Their analytic values can be easily expressed as functions of

156

the model parameters:

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

157 It is interesting to note that the expected time in the StSI model is n and does not depend
158 on the migration rate (DURRETT 2008). The variance is however, and as expected, a function
159 of both n and M . For the SSPSC model, the expected coalescence time is a function of both
160 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.
161 Indeed, when the population size change is very ancient, even if α is very large the expected
162 coalescence time will mostly depend on the present-day population size, N_0 . Similarly, when
163 T is small it will mostly depend on N_1 . The relationship that we mentioned above between
164 n and α (and between M and $1/T$) can be seen by noting that when T is close to zero
165 (and M is large), the expectations under the two models are α and n , and the variances are
166 $\text{Var}(T_2^{SSPSC}) \approx 1 + 2\alpha(\alpha - 1) - (\alpha - 1)^2 = \alpha^2$ and $\text{Var}(T_2^{StSI}) \approx n^2$. This exemplifies the
167 intuitive rationale presented above. This relationship is approximate and will be explored
168 below, but can be illustrated in more general terms by identifying scenarios with similar
169 moments.

170 As figure 2 shows, the two models provide near-identical pairs of values for $(\mathbb{E}(T_2), \text{Var}(T_2))$
171 for “well chosen” parameters (T, α) and (M, n) . Here by setting T to 0.1 (and M to 9, *i.e.*
172 $1/M \approx 0.11$) whereas α and n were allowed to vary from 1 to 100, and from 2 to 100,
173 respectively, we see that the two models exhibit very similar behaviours. We also plotted a
174 second example obtained by setting M to 0.5 and T to 1.09, and varying n and α as above.
175 These examples illustrate how n and α (respectively, M and $1/T$) are intimately related.

176 [Figure 2 about here.]

177 The near-identical values obtained for the expectation and variance under the two models

178 explains why it may be difficult to separate models of population size change from models of
179 population structure when the number of independent genetic markers is limited. However,
180 the differences between the distributions of coalescence times under the two models suggest
181 that we can probably go further and identify one model from another. For instance, figure
182 3 shows that even in cases where the first two moments are near-identical ($T = 0.1$ and
183 $\alpha = 10$ versus $M = 7$ and $n = 9$), it should be theoretically possible to distinguish them.
184 This is exactly what we aim to do in the next section. In practice, we will assume that
185 we have a sample of n_L independent T_2 values (corresponding to n_L independent *loci*) and
186 will use these T_2 values to (i) estimate the parameter values that best explain this empirical
187 distribution under the two models of interest, (ii) use a statistical test to compare the
188 empirical distribution with the expected distribution for the ML estimates and reject (or
189 not) one or both of the models. For simplicity, and to make it easier to read, we will often
190 use the term *loci* in the rest of the manuscript when we want to mention the number of
191 independent T_2 values.

192

[Figure 3 about here.]

193 Model choice and parameter estimation

194 *General principle and parameter combinations:* Given a sample (t_1, \dots, t_{n_L}) of n_L independent
195 observations of the random variable T_2 , we propose a parameter estimation procedure and
196 a goodness-of-fit test to determine whether the observed distribution of the T_2 values is
197 significantly different from that expected from the theoretical T_2^{SSPSC} or T_2^{StSI} distributions.
198 This sample can be seen as a set of T_2 values obtained or estimated from n_L independent loci.
199 We took a Maximum Likelihood (ML) approach to estimate the parameters (T, α) and (M, n)
200 under the hypothesis that the n_L -sample was generated under the T_2^{SSPSC} and the T_2^{StSI}
201 distributions, respectively (see Supplementary materials for the details of the estimation
202 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}
203 reference distributions. The Kolmogorov-Smirnov (*KS*) test which allows to compare a
204 sample with a reference distribution was then used to determine whether the observed n_L
205 sample could have been generated by the respective demographic models. In other words
206 this allowed us to reject (or not) the hypothesis that the (t_1, \dots, t_{n_L}) sample was a realization
207 of the reference distributions (T_2^{StSI} or T_2^{SSPSC}). Note that the estimation procedure and
208 the *KS* test were performed on independent sets of T_2 values. We thus simulated twice as
209 many T_2 values as needed ($2n_L$ instead of n_L). With real data that would require that half
210 of the loci be used to estimate $(\widehat{T}, \widehat{\alpha})$ and $(\widehat{M}, \widehat{n})$, whereas the other half would be used to
211 perform the *KS* test.

212 We expect that if the estimation procedure is accurate and if the *KS* test is performing
213 well we should reject the SSPSC (respectively, the StSI) model when the data were simu-
214 lated under the StSI (resp., the SSPSC) model. On the contrary we should not reject data
215 simulated under the SSPSC (resp., the StSI) model when they were indeed simulated under
216 that model. To validate our approach we used (t_1, \dots, t_{2n_L}) data sampled from the two T_2
217 distributions and quantified how the estimation procedure and the *KS* test performed. In
218 order to do that, we varied the parameter values $((T, \alpha)$ and $(M, n))$ for various $2n_L$ values

219 as follows. For T and α we used all 36 pairwise combinations between these two sets of
 220 values (0.1, 0.2, 0.5, 1, 2, 5), and (2, 4, 10, 20, 50, 100), respectively. For M and n we used
 221 all the 48 combinations between the following values (0.1, 0.2, 0.5, 1, 5, 10, 20, 50) and (2,
 222 4, 10, 20, 50, 100), respectively. For $2n_L$ we used the following values (40, 100, 200, 400,
 223 1000, 2000, 20000). Altogether we tested 588 combinations of parameters and number of
 224 loci. For each $2n_L$ value and for each parameter combination (T, α) (or (M, n)) we realized
 225 100 independent repetitions of the following process. We first simulated a sample of $2n_L$
 226 values using the *pdfs* of the SSPSC (resp. StSI) model with (T, α) (resp. (M, n)). We then
 227 used the first n_L values to obtain the ML estimates $(\widehat{T}, \widehat{\alpha})$ for the SSPSC model and $(\widehat{M}, \widehat{n})$
 228 for the StSI model. Then, we performed a *KS* test using a 0.05 threshold on the second half
 229 of the simulated data (*i.e.* n_L values) with each of the theoretical distributions defined by
 230 the estimated parameters. Finally, after having repeated this process 100 times we recorded
 231 all estimated parameters and counted the number of times we rejected the SSPSC and StSI
 232 models for each parameter combination and each $2n_L$ value.

233 *Maximum likelihood estimation in the SSPSC case:* We know from section the *pdf* of the
 234 coalescence time in the SSPSC model of two genes. We can thus write the likelihood function
 235 for any couple of parameters (α, T) , given one observation t_i as:

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

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

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

237 and taking the *log* it gives:

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

238 **Lemma 0.1** Given a set of n_L independent observations $\{t_1, t_2, \dots, t_{n_L}\}$, the log-likelihood
 239 function $\log(\mathbb{L}_{SSPSC})$ has no critical points in \mathbb{R}^2 .

240 For the proof and some comments, see Supplementary Materials.

241 As a consequence of this lemma, we take $(\hat{\alpha}, \hat{T}) = \operatorname{argmax}_{a \in \{1, \dots, n_L\}} \{\log(\mathbb{L}_{SSPSC}(m_a))\}$
 242 as the Maximum Likelihood estimates, where

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

243 with

$$K = \sum_{i=1}^{n_L} \mathbb{I}_{t_i < t_a}$$

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

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

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

RESULTS

252

253 Figure 4 shows, for various values of n_L , the results of the estimation of α (panels (a), (c), and
254 (e), for simulations assuming $\alpha = 10$ and $T = (0.1, 1, 2)$, respectively ; see Supplementary
255 Material for the other values) and the estimation of n (panels (b), (d), and (f) for simulations
256 with $n = 10$ and $M = (10, 1, 0.5)$, respectively; see Supplementary Material for the other
257 values, corresponding to 26 figures and 168 panels). The first thing to notice is that both
258 α and n are increasingly well estimated as n_L increases. This is what we expect since n_L
259 represents the amount of information (the number of T_2 values or independent loci.) The
260 second thing to note is that the two parameters are very well estimated when we use 10,000
261 values of T_2 . This is particularly obvious for n compared to α , probably because n must be
262 an integer, whereas α is allowed to vary continuously. For instance, for most simulations we
263 find the exact n value (without error) as soon as we have more than 1000 loci. However, we
264 should be careful in drawing very general rules. Indeed, when fewer T_2 values are available
265 (*i.e.* fewer independent loci), the estimation precision of both parameters depends also on
266 T and M , respectively. Interestingly, the estimation of α and n are remarkable even when
267 these parameters are small. This means that even “mild” bottlenecks may be very well
268 quantified (see for instance the Supplementary materials for $\alpha = 2$, T values between 0.1
269 and 1 when we use only 1000 loci). We should also note that when the bottleneck is very
270 old ($T = 5$) the estimation of the parameters is rather poor and only starts to be reasonable
271 and unbiased for $n_L = 10,000$. This is not surprising since the expected $T_{MRC A}$ is 1. Under
272 the SSPSC model most genes will have coalesced by $t = 5$, and should therefore exhibit T_2
273 values sampled from a stationary population (*i.e.* $\alpha = 1$). As the number of loci increases,
274 a small proportion will not have coalesced yet and will then provide information on α . The
275 expected proportion of genes that have coalesced by $T = 5$ is 0.993.

276 Figure 5 shows for various values of n_L the results of the estimation of T (panels (a), (c),
277 and (e), for simulations assuming $T = 0.2$ and $\alpha = (2, 20, 100)$, respectively; see Supplemen-
278 tary Material for the other values) and the estimation of M (panels (b), (d), and (f), for

279 simulations with $M = 20$ and $n = (2, 20, 100)$, respectively; see Supplementary Material for
280 the other values). As expected again, the estimates are getting better as n_L increases. For
281 the values shown here we can see that T , the age of the bottleneck is very well estimated
282 even when $\alpha = 2$ (for $n_L = 10,000$). In other words, even a limited bottleneck can be very
283 precisely dated. For stronger bottlenecks fewer loci (between 500 and 1000) are needed to
284 still reach a high precision. This is particularly striking given that studies suggest that it
285 is hard to identify bottlenecks with low α values (GIROD *et al.* 2011). Interestingly, the
286 panels (b), (d) and (f) seem to suggest that it may be more difficult to estimate M than
287 T . As we noted above this observation should be taken with care. Indeed, T and M are
288 not equivalent in the same way as α and n . This is why we chose to represent a value of
289 M such that $M = 1/T$, and why one should be cautious in drawing general conclusions
290 here. Altogether this and the previous figure show that it is possible to estimate with a
291 high precision the parameters of the two models by using only 500 or 1000 loci from a single
292 diploid individual. There are also parameter combinations for which much fewer loci could
293 be sufficient (between 50 and 100).

294 In Figure 6 we show some results of the KS test for the two cases (See the Supplementary
295 Materials for the other parameter combinations). In the left-hand panels ((a), (c), and (e))
296 the data were simulated under the SSPSC model and we used the StSI model as a reference
297 (*i.e.* we ask whether we can reject the hypothesis that genetic data were generated under
298 a structured model when they were actually generated under a model of population size
299 change). In the right-hand panels ((b), (d) and (f)) the same data were compared using the
300 SSPSC model as reference and we computed how often we rejected them using a 5% rejection
301 threshold. The left-hand panels exhibit several important features. The first is that, with
302 the exception of $T_2 = 5$ we were able to reject the wrong hypothesis in 100% of the cases
303 when we used 10,000 independent T_2 values.

304 This shows that our estimation procedure (as we saw above in figures 4 and 5) and the
305 KS test are very powerful. The second feature is that for $T = 5$, the test performs badly

306 whatever the number of independent loci (at least up to 10,000). This is expected since the
 307 expected $T_{MRC A}$ of two genes is $t = 1$, and 99.3% of the loci will have coalesced by $t = 5$.
 308 This means that out of the 10,000, only *c.a.* 70 loci are actually informative regarding the
 309 pre-bottleneck population size. Another important feature of the left-hand panels is that
 310 the best results are generally obtained for $T = 1, 0.5$ and 2, whichever the value of α . This
 311 is in agreement with GIROD *et al.* (2011) in that very recent population size changes are
 312 difficult to detect and quantify. The observation is valid for ancient population size changes
 313 as well. The right-hand panels are nearly identical, whichever α value we used (see also
 314 Supplementary Materials), and whichever number of T_2 values we use. They all show that
 315 the *KS* test always rejects a rather constant proportion of data sets. This proportion varies
 316 between 3 and 15%, with a global average of 8.9%. Altogether our *KS* test seems to be
 317 conservative. This is expected because for low n_L values the estimation of the parameters
 318 will tend to be poor. Since the *KS* test uses a reference distribution based on the estimated
 319 rather than the true values, it will reject the simulated data more often than the expected
 320 value of 5%.

321 Figure 7 is similar to Figure 6 but the data were simulated under the StSI model and
 322 the *KS* test was performed first using the SSPSC model as a reference ((a), (c), (e)) and
 323 then using the StSI model as a reference ((b), (d), (f)). The left-hand panels ((a), (c), and
 324 (e)) show results when we ask whether we can reject the hypothesis that genetic data were
 325 generated under a population size change model when they were actually generated under
 326 a model of population structure. In the right-hand panels ((b), (d), and (f)) we computed
 327 how often we rejected the hypothesis that genetic data were generated under the StSI model
 328 when they were indeed generated under that model of population structure. Altogether,
 329 the left-hand panels suggest that the results are generally best when $M = (0.1, 0.2, 1)$, but
 330 that we get very good results for most values of M when we have 10,000 loci and can reject
 331 the SSPSC when they were actually generated under the StSI model. The right-hand panels
 332 show, as in Figure 6, that for all the values of n_L and n we reject a rather constant proportion

333 of data sets (between 5 and 10%). Altogether the two previous figures (figures 6 and 7) show
334 that it is possible to identify the model under which genetic data were generated by using
335 genetic data from a single diploid individual.

336 Figure 8 is divided in four panels showing the relationships between T and M (panels
337 (b) and (d), for various values of α and n) and between α and n (panels (a) and (c),
338 for various values of T and M). In each of the panels we simulated data under a model for
339 specific parameter values represented on the x-axis, and estimated parameters from the other
340 model, and represented the estimated value on the y-axis. Since we were interested in the
341 relationship between parameters (not in the quality of the estimation, see above), we used the
342 largest n_L value and plotted the average of 100 independent estimation procedures. In panel
343 (a) we simulated a population size change (SSPSC) for various T values (represented each by
344 a different symbol) and several values of α on the x-axis. We then plotted the estimated value
345 of \hat{n} for each case (*i.e.* when we assume that the data were generated under the StSI model).
346 We find a striking linear relationship between these two parameters conditional on a fixed T
347 value. For instance, a population bottleneck by a factor 50 that happened N_0 generations ago
348 ($T = 1$) is equivalent to a structured population with $\hat{n} \approx 22$ islands (and $\widehat{M} \approx 0.71$). Panel
349 (c) is similar and shows how data simulated under a structured population generates specific
350 parameters of population bottlenecks. Panels (b) and (d) show the relationship between T
351 and M . We have plotted as a reference the curve corresponding to $y = 1/x$. As noted above
352 and shown on this graph, this relationship is only approximate and depends on the value
353 of α and n . Altogether, this figure exhibits the profound relationships between the model
354 parameters. They show that the qualitative relationships between α and n , and between T
355 and $1/M$ discussed above are indeed real but only correct up to a correcting factor. Still this
356 allows us to identify profound relationships between population structure and population
357 size change.

358

[Figure 4 about here.]

359

[Figure 5 about here.]

360

[Figure 6 about here.]

361

[Figure 7 about here.]

362

[Figure 8 about here.]

DISCUSSION

363

364 In this study we have analysed the distribution of coalescence times under two simple demo-
365 graphic models. We have shown that even though these demographic models are strikingly
366 different (Figure 1) there is always a way to find parameter values for which both models
367 will have the same first two moments (Figure 2). We have also shown that there are intrin-
368 sic relationships between the parameters of the two models (Figure 8). However, and this
369 is a crucial point, we showed that the distributions were different and could therefore be
370 distinguished. Using these distributions we developed a *ML* estimation procedure for the
371 parameters of both models $(\widehat{T}, \widehat{\alpha})$ and $(\widehat{M}, \widehat{n})$ and showed that the estimates are accurate,
372 given enough genetic markers. Finally, we showed that by applying a simple *KS* test we were
373 able to identify the model under which specific data sets were generated. In other words, we
374 were able to determine whether a bottleneck signal detected in a particular data set could
375 actually be caused by population structure using genetic data from a single individual. The
376 fact that a single individual provides enough information to estimate demographic param-
377 eters is in itself striking (see in particular the landmark paper by LI and DURBIN (2011)),
378 but the fact that one individual (or rather sometimes as few as 500 or 1000 loci from that
379 one individual) potentially provides us with the ability to identify the best of two (or more)
380 models is even more remarkable. The PSMC (pairwise sequentially markovian coalescent)
381 method developed by LI and DURBIN (2011) reconstructs a theoretical demographic history
382 characterized by population size changes, assuming a single non structured population. Our
383 study goes further and shows that it could be possible to test whether the signal identified
384 by the PSMC is due to actual population size changes or to population structure. However,
385 models putting together these two scenarios have being proposed. In WAKELEY (1999), a
386 model considering a structured population who went through a bottleneck in the past is
387 developed, allowing to estimate the *time* and the *ratio* of the bottleneck. Moreover, WAKE-
388 LEY (1999) discussed the idea that, in structured populations, changes in the migration rate
389 and/or the number of islands (demes) can shape the observed data in the same way that

390 effective population size changes do. Hence, we think that our work could be helpful to the
391 aim of setting these two scenarios apart in order to detect (for example) false bottleneck
392 signals. Nevertheless, while our study provides several new results, there are still several
393 important issues that need to be discussed and much progress that can still be made.

394 T_2 and molecular data

395 The first thing to note is that we assume, throughout our study, that we have access to the
396 coalescence times T_2 . In real data sets, this is never the case and the T_2 are rarely estimated
397 from molecular data. While this may seem as a limitation, we should note that the recent
398 method of LI and DURBIN (2011), that uses the genome sequence of a single individual to
399 infer the demographic history of the population it was sampled from, actually estimates the
400 distribution of T_2 values. In its current implementation the PSMC software does not output
401 this distribution but it could be modified to do it. Note however, that the PSMC should
402 only be able to provides a discretized distribution in the form of a histogram with classes
403 defined by the number of time periods for which population size estimates are computed.
404 In any case, this suggests that it is in theory possible to use the theoretical work of Li and
405 Durbin to generate T_2 distributions, which could then be used with our general approach.
406 Moreover, it should be possible to use the theory developed here to compute, conditional
407 of the T_2 distribution, the distribution of several measures of molecular polymorphism. For
408 instance, under an infinite site mutation model it is in principle possible to compute the
409 distribution of the number of differences between pairs of non recombining sequences for
410 the two demographic models analysed here. Similarly, assuming a single stepwise mutation
411 model it should be possible to compute the distribution of the number of repeat differences
412 between two alleles conditional on T_2 . However, we must add here that while it is easy to use
413 these distributions to simulate genetic data (rather than T_2 values) it is not straightforward
414 to then use the genetic data to estimate the models' parameters and apply a test to identify

415 the model under which the data were simulated. This would probably require a Khi-2 test
416 since discrete rather continuous distributions would be compared. This is an issue that would
417 deserve a full and independent study.

418 **Demographic models**

419 In our study we limited ourselves to two simple but widely used models. It would thus be
420 important to determine the extent to which our approach could apply to other demographic
421 models. The n-island or StSI model is a widely used model and it was justified to use it
422 here. One of its strongest assumptions is that migration is identical between all demes.
423 This is likely to be problematic for species with limited vagility. In fact, for many species a
424 model where migration occurs between neighbouring populations such as the stepping-stone
425 is going to be more likely. At this stage it is unclear whether one could derive analytically
426 the *pdf* of T_2 for a stepping-stone model. The work by HERBOTS (1994) suggests that it
427 may be possible to compute it numerically by inverting the Laplace transform derived by
428 this author. This work has not been done to our knowledge and would still need to be done.
429 Interestingly, this author has also shown that it is in principle possible to derive analytically
430 the *pdf* of T_2 in the case of a two-island model with populations of different sizes. Again,
431 this still needs to be done.

432 The SSPSC model has also been widely used (ROGERS and HARPENDING 1992) and
433 represents a first step towards using more complex models of population size change. For
434 instance, the widely used method of BEAUMONT (1999) to detect, date and quantify pop-
435 ulation size changes (GOOSSENS *et al.* 2006; QUÉMÉRÉ *et al.* 2012; SALMONA *et al.* 2012)
436 assumes either an exponential or a linear population size change. It should be straightfor-
437 ward to compute the *pdf* of T_2 under these two models because, as we explained above, the
438 coalescent theory for populations with variable size has been very well studied (DONNELLY
439 and TAVARE 1995; TAVARÉ 2004) and it is possible to write the *pdf* of T_2 for any demo-

440 graphic history involving any type of population size changes. To go even further one could
441 in principle use the history reconstructed by the PSMC (LI and DURBIN 2011) as the ref-
442 erence model of population size change and compare that particular demographic history to
443 an n-island model using our results, and our general approach. At the same time, we should
444 note that for complex models of population size change, including relatively simple ones such
445 as the exponential model of BEAUMONT (1999), it is not straightforward to compute the
446 number of differences between non recombining sequences. Significant work would probably
447 be needed to apply the general approach outlined in our study to specific demographic mod-
448 els. But we believe that the possibilities opened by our study are rather wide and should
449 provide our community with new interesting problems to solve in the next few years.

450 **Comparison with previous work and generality our of results**

451 The present work is part of a set of studies aimed at understanding how population
452 structure can be mistaken for population size change and at determining whether studies
453 identifying population size change are mistaken or valid (CHIKHI *et al.* 2010; HELLER *et al.*
454 2013; PAZ-VINAS *et al.* 2013). It is also part of a wider set of studies that have recognised in
455 the last decade the importance of population structure as potential factor biasing inference
456 of demographic (LEBLOIS *et al.* 2006; STÄDLER *et al.* 2009; PETER *et al.* 2010; CHIKHI
457 *et al.* 2010; HELLER *et al.* 2013; PAZ-VINAS *et al.* 2013) or selective processes (CURRAT
458 *et al.* 2006). Here we demonstrated that it is indeed possible to separate the SSPSC and
459 StSI models. While we believe that it is an important result, we also want to stress that
460 we should be cautious before extending these results to any set of models, particularly
461 given that we only use the information from T_2 . Much work is still needed to devise new
462 tests and estimation procedures for a wider set of demographic models and using more
463 genomic information, including recombination patterns as in the PSMC method (LI and
464 DURBIN 2011). Beyond the general approach outlined here we would like to mention the

465 study of PETER *et al.* (2010) who also managed to separate one structure and one PSC
466 (*Population Size Change*) model. These authors used an ABC approach to separate a model
467 of exponential PSC from a model of population structure similar to the StSI model. Their
468 structured model differs from ours by the fact that it is not an equilibrium model. They
469 assumed that the population was behaving like an n-island model in the recent past, until T
470 generations in the past, but that before that time, the ancestral population from which all
471 the demes derived was not structured. When T is very large their model is identical to the
472 StSI, but otherwise it may be quite different. The fact that they managed to separate the two
473 models using an ABC approach is promising as it suggests that there is indeed information
474 in the genetic data for models beyond those that we studied here. We can therefore expect
475 that our approach may be applied to a wider set of models. We should also add that in their
476 study these authors use a much larger sample size (25 diploid individuals corresponding to 50
477 genes). They used a maximum of 200 microsatellites which corresponds therefore to 10,000
478 genotypes, a number very close to the maximum number we used here. Altogether our study
479 provides new results and opens up new avenues of research for the distribution of coalescent
480 times under complex models.

481 **Sampling and population expansions**

482 Recent years have also seen an increasing recognition of the fact that the sampling scheme
483 together with population structure may significantly influence demographic inference (WAKELEY
484 1999; STÄDLER *et al.* 2009; CHIKHI *et al.* 2010; HELLER *et al.* 2013). For instance, WAKE-
485 LEY (1999) showed that in the n-island model genes sampled in different demes will exhibit
486 a genealogical tree similar to that expected under a stationary Wright-Fisher model. Since
487 our work was focused on T_2 we mostly presented our results under the assumption that the
488 two genes of interest were sampled in the same deme. For diploids this is of course a most
489 reasonable assumption. However, we should note that the results presented above also allow

490 us to express the distribution of T_2 when the genes are sampled in different demes. We did
491 not explore this issue further here, but it would be important to study the results under such
492 conditions. Interestingly we find that if we assume that the two genes are sampled in two
493 distinct demes, we can detect population expansions rather than bottlenecks. This could
494 happen if we considered a diploid individual whose parents came from different demes. In
495 that case, considering the two genes sampled in the deme where the individual was sampled
496 would be similar to sampling his two parental genes in two different demes. Interestingly,
497 this has also been described by PETER *et al.* (2010) who found that when the 25 individ-
498 uals were sampled in different demes, they would detect population size expansions rather
499 than bottlenecks. Our results are therefore in agreement with theirs. Similarly HELLER
500 *et al.* (2013) also found that some signals of population expansion could be detected under
501 scattered sampling schemes.

502 **Conclusion: islands within individuals**

503 To conclude, our results provide a general framework that can in theory be applied to whole
504 families of models. We showed for the first time that genomic data from a single individual
505 could be used to estimate parameters that have to our knowledge never been estimated. In
506 particular we showed that we were able to estimate the number of islands (and the number
507 of migrants) in the StSI model. This means that one can in principle use genomic data from
508 non model organisms to determine how many islands make up the metapopulation from
509 which one single individual was sampled. This is of course true as well for model organisms
510 but it is particularly meaningful for species for which the number of individuals with genomic
511 data is limited. Our ability to estimate n is one of the most powerful results of our study.
512 While such estimates should not be taken at face value, they surely should be obtained across
513 species for comparative analyses. Also, during the last decade there has been a major effort
514 to use programs such as STRUCTURE (PRITCHARD *et al.* 2000) to estimate the number of

515 "subpopulations" within a particular sample. Our work suggests that we might in principle
516 provide additional results with only one individual. It is important to stress though that the
517 answer provided here is very different from those obtained with STRUCTURE and similar
518 methods and programs (PRITCHARD *et al.* 2000; GUILLOT *et al.* 2005; CHEN *et al.* 2007;
519 CORANDER *et al.* 2004). We do not aim at identifying the populations from which a set
520 of individuals come. Rather we show that his/her genome informs us on the whole set of
521 populations. In other words, even though we assume that there are n populations linked by
522 gene flow, we show that each individual, is somehow a genomic patchwork from this (poorly)
523 sampled metapopulation. We find these results reassuring, in an era where genomic data are
524 used to confine individuals to one population and where division rather than connectivity is
525 stressed.

ACKNOWLEDGEMENTS

526

527 We are grateful to S. Boitard and S. Grusea for numerous and fruitful discussions and
528 input throughout the development of this work. We thank I. Paz, I. Pais, J. Salmona,
529 J. Chave for discussion and helpful comments that improved the clarity of the text. We
530 thank J. Howard and the Fundação Calouste Gulbenkian for their continuous support. We
531 are grateful to the genotoul bioinformatics platform Toulouse Midi-Pyrenees for providing
532 computing and storage resources. This work was partly performed using HPC resources from
533 CALMIP (Grant 2012 - projects 43 and 44) from Toulouse, France. This study was funded
534 by the Fundação para a Ciência e Tecnologia (ref. PTDC/BIA- BIC/4476/2012), the Projets
535 Exploratoires Pluridisciplinaires (PEPS 2012 Bio-Maths-Info) project, the LABEX entitled
536 TULIP (ANR-10-LABX-41) as well as the Pôle de Recherche et d'Enseignement Supérieur
537 (PRES) and the Région Midi-Pyrénées, France.

LITERATURE CITED

538

539 BEAUMONT, M. A., 1999 Detecting population expansion and decline using microsatellites.
540 *Genetics* *153*(4): 2013–2029.

541 BEAUMONT, M. A., 2010 Approximate Bayesian computation in evolution and ecology.
542 *Annual review of ecology, evolution, and systematics* **41**: 379–406.

543 BEAUMONT, M. A., W. ZHANG, and D. J. BALDING, 2002 Approximate Bayesian com-
544 putation in population genetics. *Genetics* *162*(4): 2025–2035.

545 BROQUET, T., S. ANGELONE, J. JAQUIERY, P. JOLY, J.-P. LENA, T. LENGAGNE,
546 S. PLENET, E. LUQUET, and N. PERRIN, 2010 Genetic bottlenecks driven by pop-
547 ulation disconnection. *Conservation Biology* *24*(6): 1596–1605.

548 CHEN, C., E. DURAND, F. FORBES, and O. FRANÇOIS, 2007 Bayesian clustering algo-
549 rithms ascertaining spatial population structure: a new computer program and a com-
550 parison study. *Molecular Ecology Notes* *7*(5): 747–756.

551 CHIKHI, L., V. C. SOUSA, P. LUISI, B. GOOSSENS, and M. A. BEAUMONT, 2010 The
552 confounding effects of population structure, genetic diversity and the sampling scheme
553 on the detection and quantification of population size changes. *Genetics* *186*(3): 983–995.

554 CORANDER, J., P. WALDMANN, P. MARTTINEN, and M. J. SILLANPÄÄ, 2004 BAPS
555 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformat-
556 ics* *20*(15): 2363–2369.

557 CORNUET, J.-M., F. SANTOS, M. A. BEAUMONT, C. P. ROBERT, J.-M. MARIN, D. J.
558 BALDING, T. GUILLEMAUD, and A. ESTOUP, 2008 Inferring population history with
559 DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformat-
560 ics* *24*(23): 2713–2719.

561 CURRAT, M., L. EXCOFFIER, W. MADDISON, S. P. OTTO, N. RAY, M. C. WHITLOCK,
562 and S. YEAMAN, 2006 Comment on “Ongoing adaptive evolution of ASPM, a brain

563 size determinant in *Homo sapiens*” and “Microcephalin, a gene regulating brain size,
564 continues to evolve adaptively in humans”. *Science* 313(5784): 172a–172a.

565 DONNELLY, P. and S. TAVARE, 1995 Coalescents and genealogical structure under neu-
566 trality. *Annual review of genetics* 29(1): 401–421.

567 DURRETT, R., 2008 *Probability models for DNA sequence evolution*. Springer.

568 GIROD, C., R. VITALIS, R. LEBLOIS, and H. FRÉVILLE, 2011 Inferring Population Decline
569 and Expansion From Microsatellite Data: A Simulation-Based Evaluation of the Msva
570 Method. *Genetics* 188(1): 165–179.

571 GOOSSENS, B., L. CHIKHI, M. ANCRENAZ, I. LACKMAN-ANCRENAZ, P. ANDAU, and
572 M. W. BRUFORD, 2006 Genetic signature of anthropogenic population collapse in orang-
573 utans. *PLoS Biology* 4(2): e25.

574 GUILLOT, G., F. MORTIER, and A. ESTOUP, 2005 GENELAND: a computer package for
575 landscape genetics. *Molecular Ecology Notes* 5(3): 712–715.

576 HELLER, R., L. CHIKHI, and H. R. SIEGISMUND, 2013 The confounding effect of popula-
577 tion structure on Bayesian skyline plot inferences of demographic history. *PLoS One* 8(5):
578 e62992.

579 HERBOTS, H. M. J. D., 1994 Stochastic models in population genetics: genealogy and
580 genetic differentiation in structured populations. Ph. D. thesis.

581 JONES, E., T. OLIPHANT, P. PETERSON, and OTHERS, 2001 SciPy: Open source scientific
582 tools for Python. [Online; accessed 2014-11-18].

583 LEBLOIS, R., A. ESTOUP, and R. STREIFF, 2006 Genetics of recent habitat contraction
584 and reduction in population size: does isolation by distance matter? *Molecular Ecol-
585 ogy* 15(12): 3601–3615.

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

- 588 NELDER, J. A. and R. MEAD, 1965 A simplex method for function minimization. The
589 computer journal 7(4): 308–313.
- 590 NIELSEN, R. and M. A. BEAUMONT, 2009 Statistical inferences in phylogeography. Molec-
591 ular Ecology 18(6): 1034–1047.
- 592 PAZ-VINAS, I., E. QUÉMÉRÉ, L. CHIKHI, G. LOOT, and S. BLANCHET, 2013 The demo-
593 graphic history of populations experiencing asymmetric gene flow: combining simulated
594 and empirical data. Molecular ecology 22(12): 3279–3291.
- 595 PETER, B. M., D. WEGMANN, and L. EXCOFFIER, 2010 Distinguishing between pop-
596 ulation bottleneck and population subdivision by a Bayesian model choice procedure.
597 Molecular Ecology 19(21): 4648–4660.
- 598 PRITCHARD, J. K., M. STEPHENS, and P. DONNELLY, 2000 Inference of population
599 structure using multilocus genotype data. Genetics 155(2): 945–959.
- 600 QUÉMÉRÉ, E., X. AMELOT, J. PIERSON, B. CROUAU-ROY, and L. CHIKHI, 2012 Genetic
601 data suggest a natural prehuman origin of open habitats in northern Madagascar and
602 question the deforestation narrative in this region. Proceedings of the National Academy
603 of Sciences 109(32): 13028–13033.
- 604 ROGERS, A. R. and H. HARPENDING, 1992 Population growth makes waves in the distri-
605 bution of pairwise genetic differences. Molecular biology and evolution 9(3): 552–569.
- 606 SALMONA, J., M. SALAMOLARD, D. FOUILLOT, T. GHESTEMME, J. LAROSE, J.-F. CEN-
607 TON, V. SOUSA, D. A. DAWSON, C. THEBAUD, and L. CHIKHI, 2012 Signature of a
608 pre-human population decline in the critically endangered Reunion Island endemic forest
609 bird *Coracina newtoni*. PloS one 7(8): e43524.
- 610 SOUSA, V. C., M. A. BEAUMONT, P. FERNANDES, M. M. COELHO, and L. CHIKHI,
611 2012, May) Population divergence with or without admixture: selecting models using an
612 ABC approach. Heredity 108(5): 521–530.

- 613 STÄDLER, T., B. HAUBOLD, C. MERINO, W. STEPHAN, and P. PFAFFELHUBER, 2009 The
614 impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided
615 populations. *Genetics* **182**(1): 205–216.
- 616 TAVARÉ, S., 2004 Part I: Ancestral inference in population genetics. In *Lectures on proba-*
617 *bility theory and statistics*, pp. 1–188. Springer.
- 618 VITTI, J. J., S. R. GROSSMAN, and P. C. SABETI, 2013 Detecting natural selection in
619 genomic data. *Annual review of genetics* **47**: 97–120.
- 620 WAKELEY, J., 1999 Nonequilibrium migration in human history. *Genetics* **153**(4): 1863–
621 1871.
- 622 WRIGHT, S., 1931 Evolution in Mendelian populations. *Genetics* **16**(2): 97.

List of Figures

623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663

- 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 $t = T$, the population size changes instantaneously from N_1 to N_0 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 $5N$ 36
- 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. 37
- 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 $E(T_2^{SSPSC}) = 9.994$, and $E(T_2^{StSI}) = 10$, $Var(T_2^{SSPSC}) = 118.7$ and $Var(T_2^{StSI}) = 118.0$. 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 $E(T_2^{SSPSC}) = 42.997$, and $E(T_2^{StSI}) = 43$, $Var(T_2^{SSPSC}) = 8905$ and $Var(T_2^{StSI}) = 8905$ 38
- 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 $M = (10, 1, 0.5)$ 39
- 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 $M = 5$ 40
- 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 41

664	7	Proportion of rejected data sets simulated under the StSI model. Panels (a),	
665		(c), and (e): the reference model is the SSPSC. Panels (b), (d), and (f): the	
666		reference model is the StSI model, <i>i.e.</i> the model under which the data were	
667		simulated. Note that for the abscissa we used $2n_L$ instead of n_L because in	
668		order to perform the <i>KS</i> test it is necessary to first estimate the parameters	
669		using n_L loci and then an independent set of n_L values of T_2	42
670	8	Relationships between parameters of the models	43

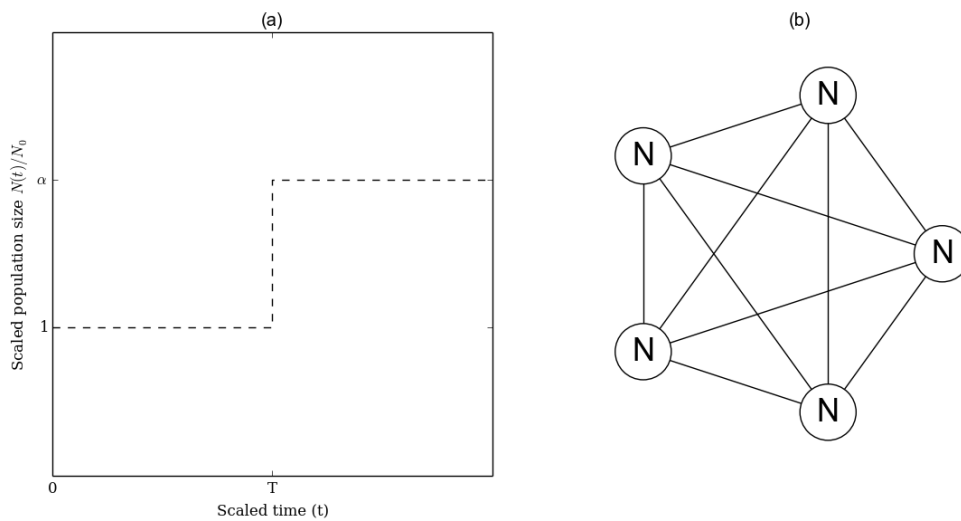


Figure 1: Demographic models. (a): Single step population size change (SSPSC) model. The x-axis represents t , the time to the past in units of generations scaled by the number of genes. At time $t = T$, the population size changes instantaneously from N_1 to N_0 by a factor α . The y-axis represents 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 $5N$.

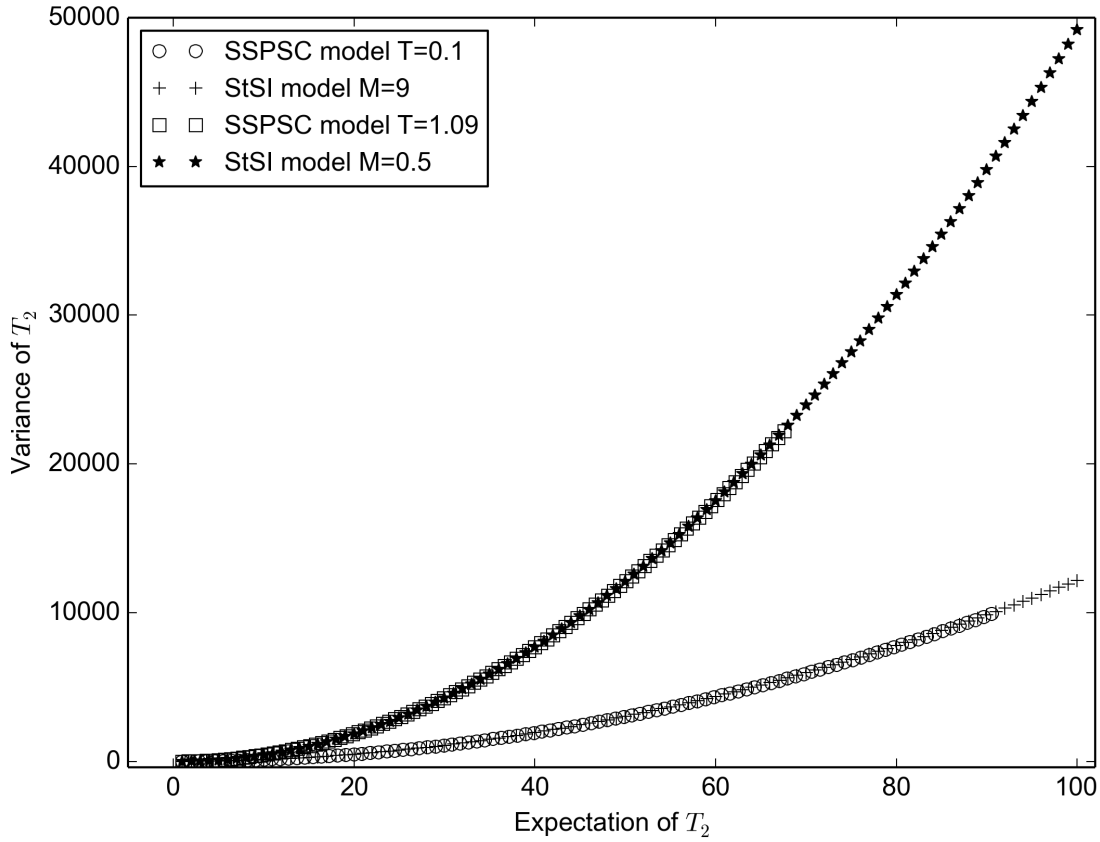


Figure 2: Expected value and Variance of T_2 under the SSPSC and StSI models. This figure illustrates how both models can have the same pair of values $(E(T_2), Var(T_2))$ for many sets of parameters. For the SSPSC model the time at which the population size change occurred was fixed to $T = 0.1$ whereas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

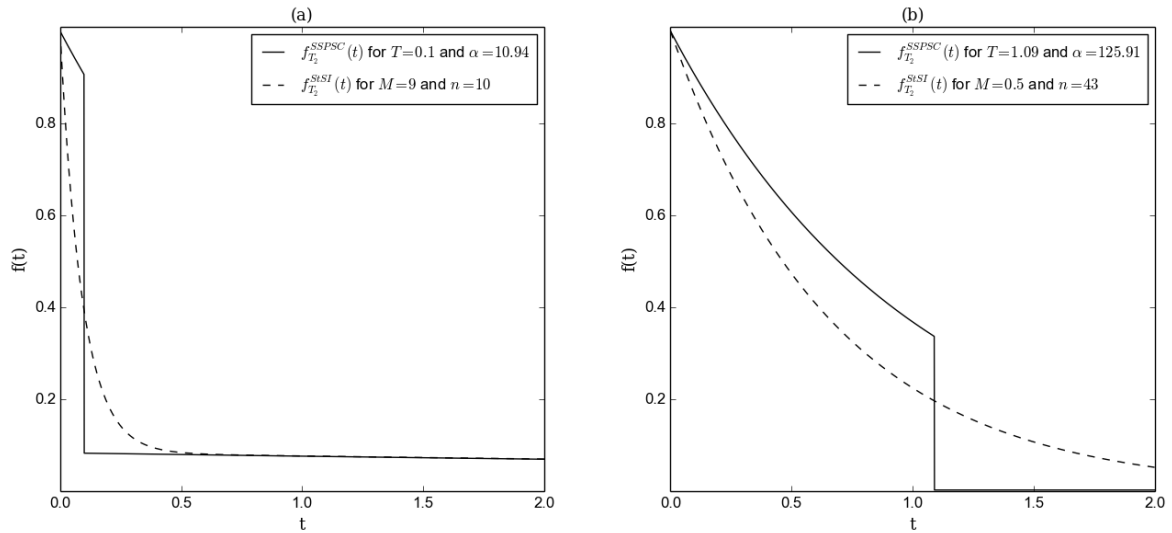


Figure 3: Density of T_2 under the SSPSC and StSI models. Two sets of parameter values (panels (a) and (b), respectively) were chosen on the basis that expectations and variances were close. Panel (a): Density for the SSPSC model with $T = 0.1$ and $\alpha = 10.94$, and for the StSI model with $M = 9$ and $n = 10$. For this set of parameters we have $E(T_2^{SSPSC}) = 9.994$, and $E(T_2^{StSI}) = 10$, $Var(T_2^{SSPSC}) = 118.7$ and $Var(T_2^{StSI}) = 118.0$. 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 $E(T_2^{SSPSC}) = 42.997$, and $E(T_2^{StSI}) = 43$, $Var(T_2^{SSPSC}) = 8905$ and $Var(T_2^{StSI}) = 8905$.

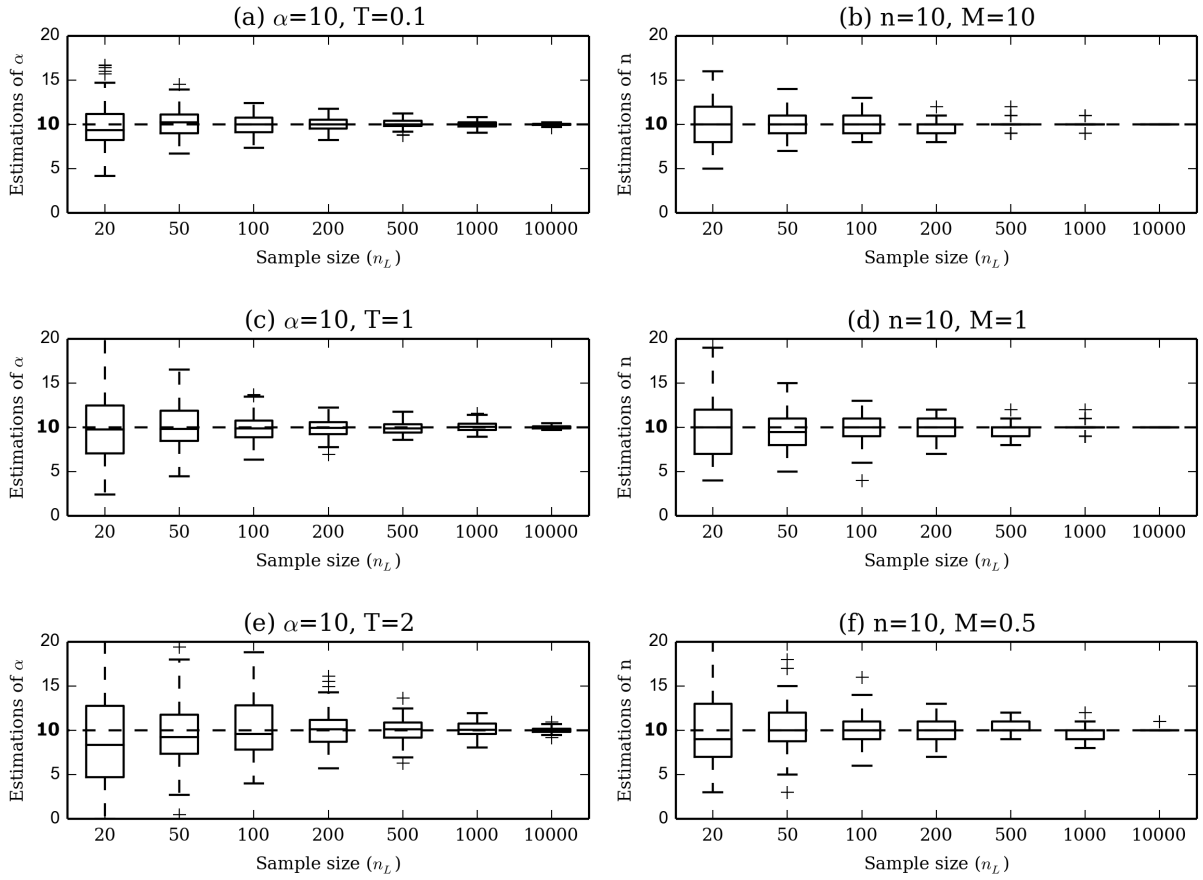


Figure 4: Estimation of α and n . Panels (a), (c) and (e): Estimation of α under the SSPSC model for different sample sizes and T values. Simulations performed with $\alpha = 10$ and $T = (0.1, 1, 2)$. Panels (b), (d) and (f): Estimation of n under the StSI model for different sample sizes and M values.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

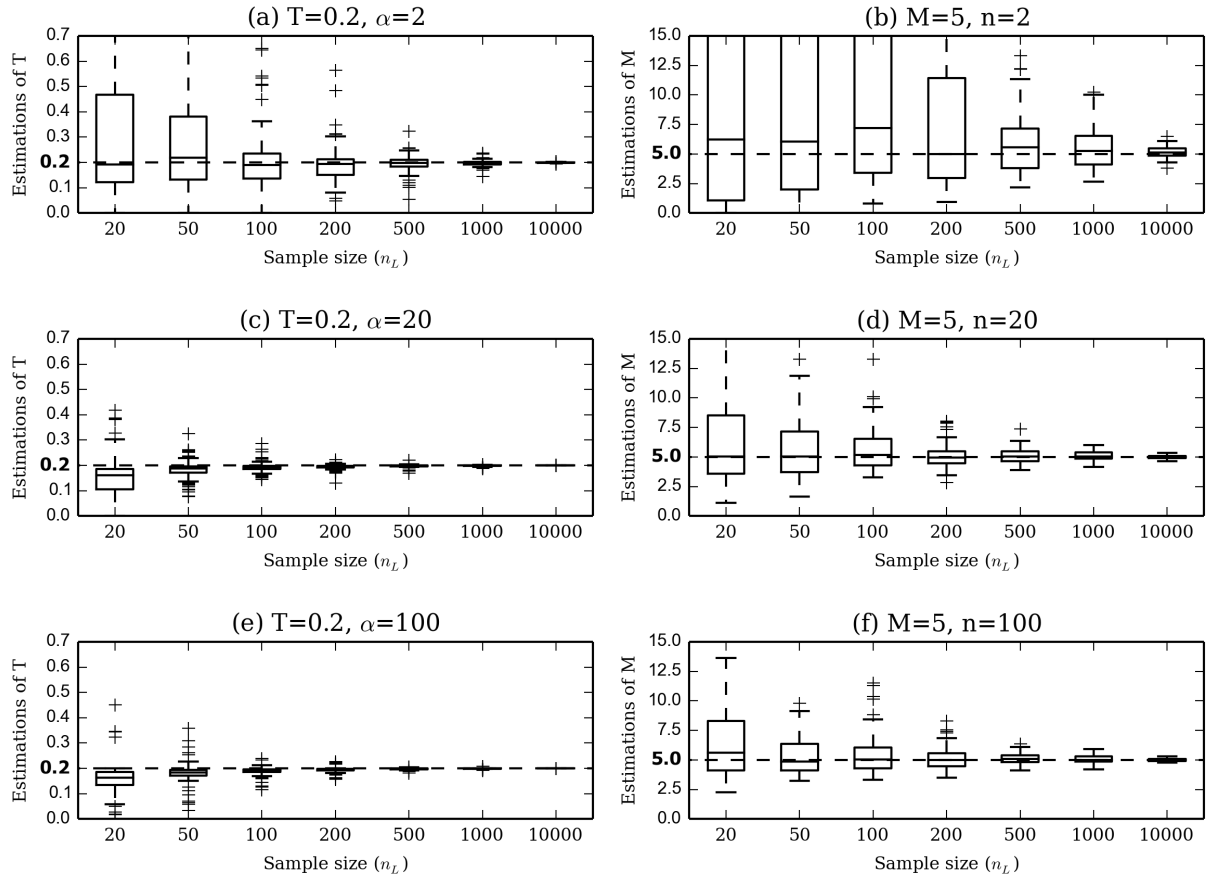


Figure 5: Estimation of T and M . Panels (a), (c), (e): Estimation of T under the SSPSC model for different sample sizes and values of α . Simulations performed with $\alpha = (2, 20, 100)$ and $T = 0.2$. Panels (b), (d), (f): Estimation of M under the StSI model for different sample sizes and values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

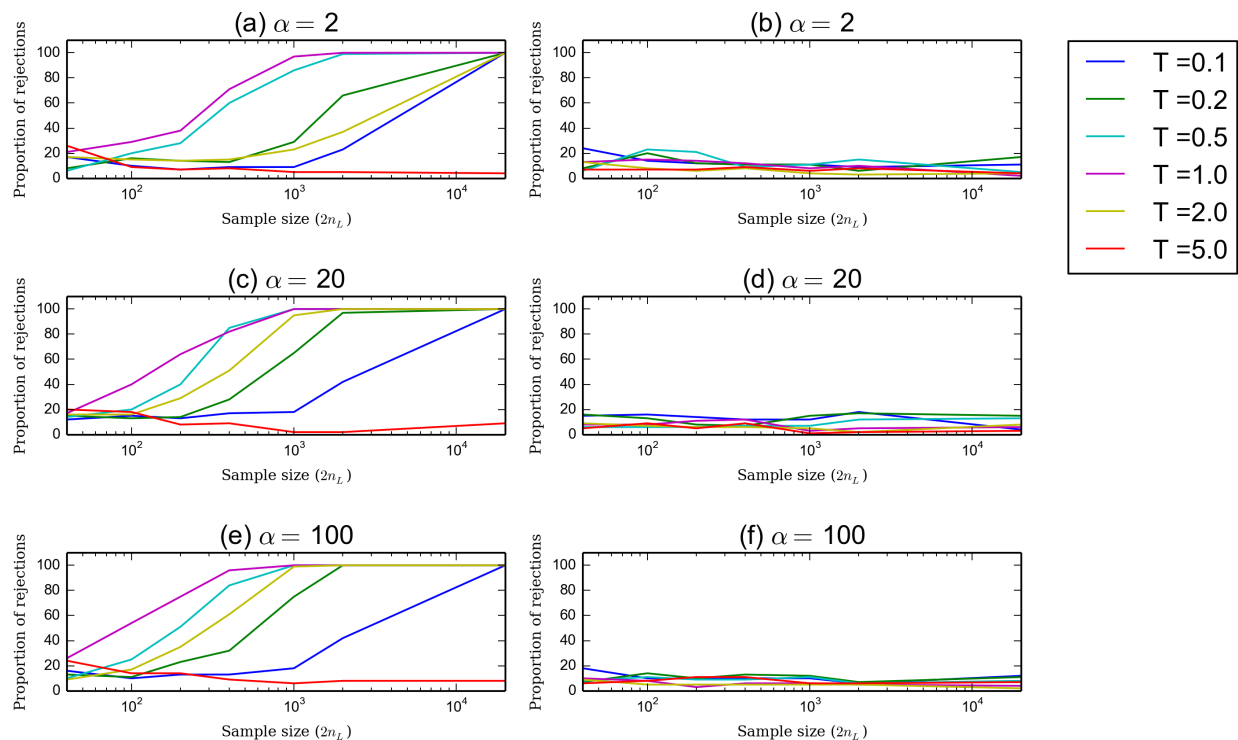


Figure 6: Proportion of rejected data sets simulated under the SSPSC model. Panels (a), (c) and (e): the reference model is the StSI model. Panels (b), (d), and (f): the reference model is the SSPSC, *i.e.* the model under which the data were simulated. Note that for the abscissa we used $2n_L$ instead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

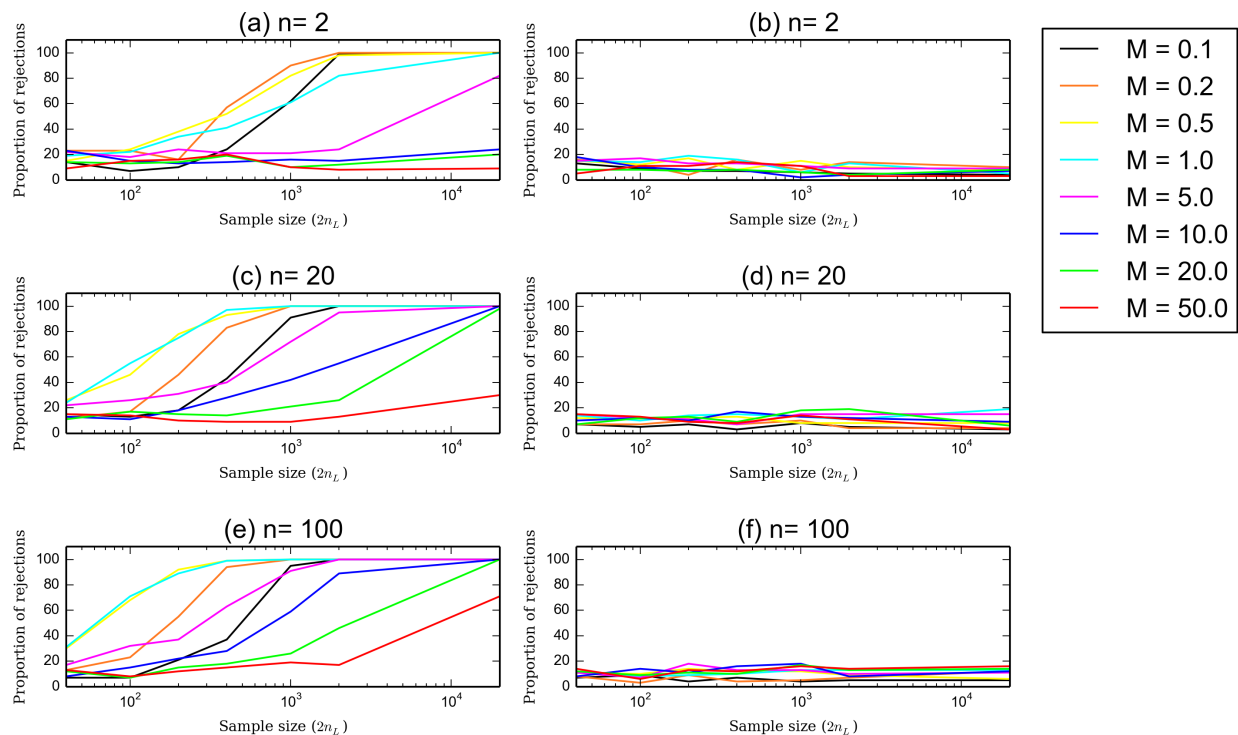


Figure 7: Proportion of rejected data sets simulated under the StSI model. Panels (a), (c), and (e): the reference model is the SSPSC. Panels (b), (d), and (f): the reference model is the StSI model, *i.e.* the model under which the data were simulated. Note that for the abscissa we used $2n_L$ instead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

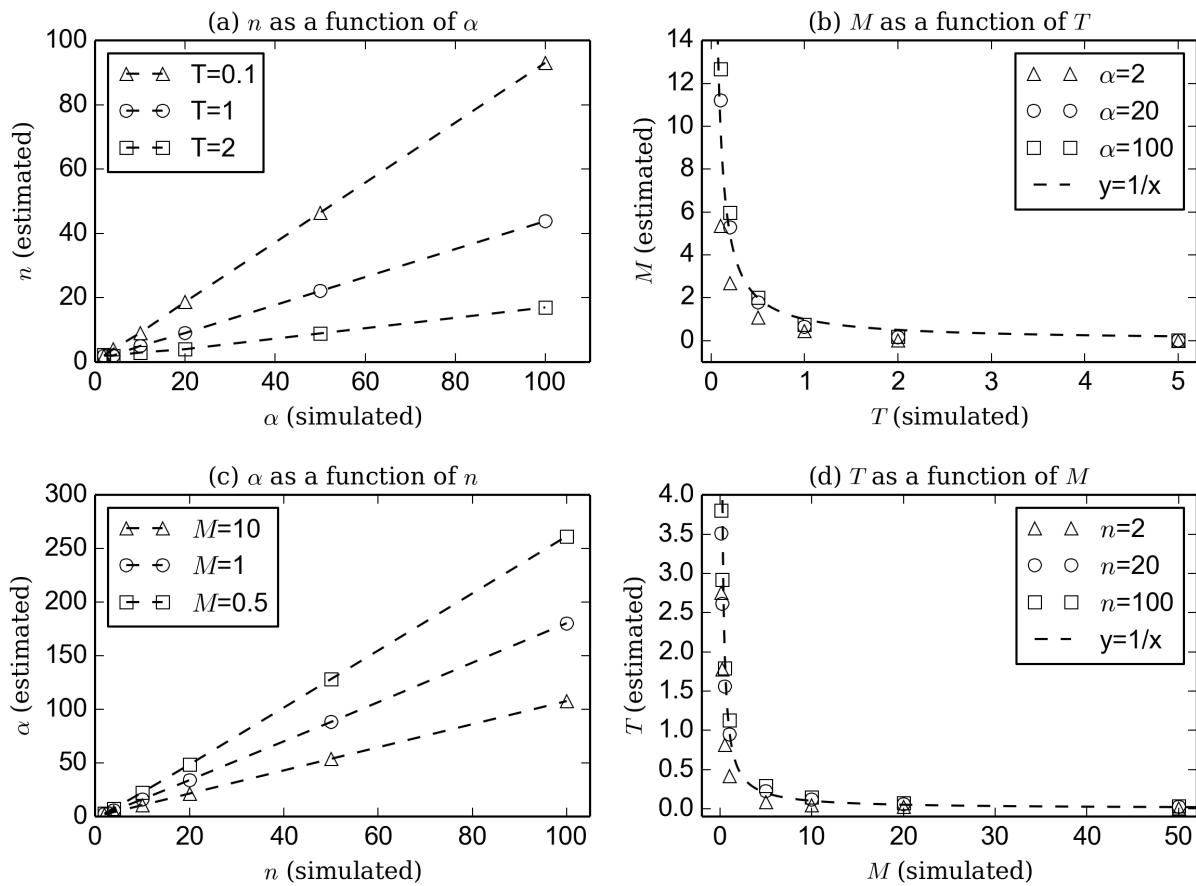


Figure 8: Relationships between parameters of the models