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.

© 2015 Elsevier Inc. All rights reserved.

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/© 2015 Elsevier Inc. All rights reserved.

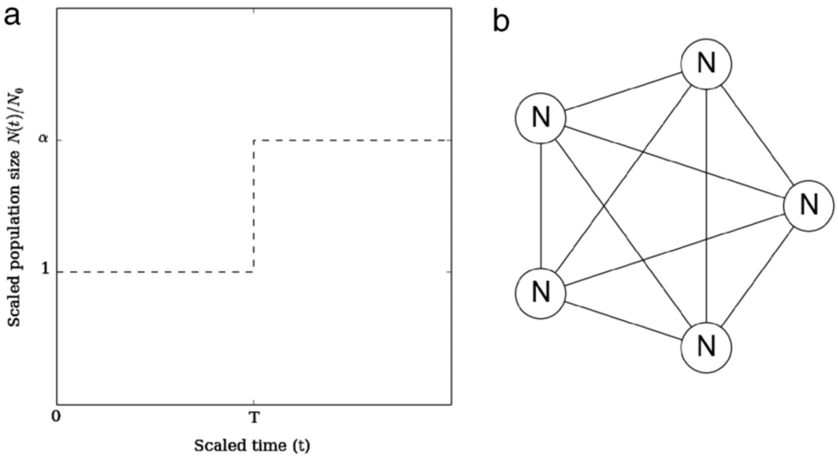

Fig. 1. Demographic models. (a) Single step population size change (SSPSC) model. The x-axis represents t, the time to the past in units of generations scaled by the number of genes. At time $t = T$, (going from the present to the past) the population size changes instantaneously from N0 to N1 by a factor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

Download English Version:

https://daneshyari.com/en/article/4502318

Download Persian Version:

https://daneshyari.com/article/4502318

Daneshyari.com