Theoretical Population Biology xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Theoretical Population Biology

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

On the decidability of population size histories from finite allele frequency spectra

Soheil Baharian, Simon Gravel *

Department of Human Genetics, McGill University, Montreal, QC, Canada
McGill University and Genome Quebec Innovation Centre, Montreal, QC, Canada

ARTICLE INFO

Article history:
Received 14 September 2017
Available online xxxx

ABSTRACT

Understanding the historical events that shaped current genomic diversity has applications in historical, biological, and medical research. However, the amount of historical information that can be inferred from genetic data is finite, which leads to an identifiability problem. For example, different historical processes can lead to identical distribution of allele frequencies. This identifiability issue casts a shadow of uncertainty over the results of any study which uses the frequency spectrum to infer past demography. It has been argued that imposing mild 'reasonableness' constraints on demographic histories can enable unique reconstruction, at least in an idealized setting where the length of the genome is nearly infinite. Here, we discuss this problem for finite sample size and genome length. Using the diffusion approximation, we obtain bounds on likelihood differences between similar demographic histories, and use them to construct pairs of very different reasonable histories that produce almost-identical frequency distributions. The finite-genome problem therefore remains poorly determined even among reasonable histories, where fits to few-parameter models produce narrow parameter confidence intervals, large uncertainties lurk hidden by model assumption.

1. Introduction

Genetic variation across individuals contains information about the evolutionary and demographic history of populations. A simple and efficient summary statistic of genomic variation commonly used in inference studies of population demography is the allele frequency spectrum, describing the proportion of segregating sites as a function of the population frequency of the derived allele (Griffiths, 2003; Marth et al., 2004; Gutenkunst et al., 2009; Lukić et al., 2011; Kim et al., 2014; Terhorst and Song, 2015). Several computational models have been proposed to reconstruct historical population sizes that are consistent with observed allele frequency spectra (Voight et al., 2005; Pickrell and Pritchard, 2012; Lukić and Hey, 2012; Excoffier et al., 2013; Gravel et al., 2013; Gao and Keinan, 2015; Jouganous et al., 2017). Under the assumption of a neutral Wright–Fisher model, these inferred histories are often taken to be representative of the effective historical population sizes (Gravel et al., 2011; Tennessen et al., 2012; Keinan and Clark, 2012; Gravel et al., 2013; Gazave et al., 2014). They are also used as baseline models to identify regions under selection (Williamson et al., 2005; Lohmueller et al., 2008; Ronen et al., 2013) and to predict patterns of deleterious variation in human genomes (Simons et al., 2014; Do et al., 2015; Gravel, 2016).

However, Myers et al. (2008) showed that the solution to this inference problem is not unique. To illustrate this, they constructed a family of distinct demographic histories whose frequency spectra under neutral Wright–Fisher evolution are identical for any sample size. This poses a serious practical challenge, since a demographic model that fits well the observed neutral diversity is not guaranteed to be historically accurate or to provide an appropriate model for deleterious variation.

On the other hand, Bhaskar and Song (2014) have argued that the families of demographic models constructed by Myers et al. are not biologically realistic, because they require historical population sizes that oscillate on arbitrarily short time-scales. They proved that we can uniquely reconstruct the underlying demography from the allele frequency spectrum if (i) we limit our search to historical population sizes that are piecewise-continuous functions of time with a given maximum number of oscillations, (ii) we have enough samples, and (iii) we assume an infinitely long genome (Bhaskar and Song, 2014).

(Myers et al., 2008; Bhaskar and Song, 2014) are both correct, but they send different messages regarding the reliability of inferred histories. Does the problem of identifiability raised by Myers et al. (2008) bear on applied inference, or should it be considered

* Corresponding author at: McGill University and Genome Quebec Innovation Centre, Montreal, QC, Canada.
E-mail address: simon.gravel@mcgill.ca (S. Gravel).

https://doi.org/10.1016/j.tpb.2017.12.008
0040-5809/

a purely theoretical result about a class of pathological functions with little biological relevance?

In this article, we seek to resolve this question by addressing the identifiability problem in more realistic scenarios where both sample size and genome length are finite. Recent work by Terhorst and Song (2015) has started to address this question by providing strict bounds on the accuracy of demographic inference based on the allele frequency spectrum. In particular, they have focused on the possibility of reconstructing history prior to a bottleneck, which is challenging because of the lost diversity (and thus lost information) during the bottleneck.

Extending this work to situations with arbitrary demographies, we argue that the problem remains poorly determined, even without bottlenecks, in the sense that vastly different population histories can produce statistically indistinguishable allele frequency spectra. Ancient history differences are most difficult to detect, as expected, but we also explain how the approach of Myers et al. can be modified to construct well-behaved, practically indistinguishable histories with somewhat more recent differences.

Our arguments are based on two simple observations. First, similar histories should produce similar frequency spectra. Second, the Myers et al. family of functions may exhibit infinitely fast oscillations, but, given the extremely small amplitude of these oscillations, they can be replaced by smooth, non-oscillating functions with tiny effect on the frequency spectrum. Thus macroscopically different histories can produce microscopic differences in the frequency distribution. These small differences could in principle be detected given an infinitely long genome, as per the Bhaskar and Song result, but they could not be detected given a finite genome of realistic length. To prove this, we first produce upper bounds on the differences between frequency spectra produced by two similar demographic models, and on the likelihood ratio between the two models given an observed frequency spectrum. Using these bounds and the family of functions given by Myers et al., we construct a family of plausible demographic histories that are very distinct but practically indistinguishable.

The main practical message from this study is that any demographic inference study based on the frequency spectrum *must* have large zones of uncertainty. These can be detected, in principle, by exploring the likelihood surface over the space of all possible functions. However, most inference studies use few-parameter demographic models and estimate parameter uncertainties assuming that these models are correct. In such studies, an excellent fit with small parameter uncertainties can still mask a model that is completely wrong.

This paper is organized as follows. We present an intuitive discussion regarding identifiable demographies in Section 2, discussing the properties of the construction of Myers et al. (2008). In Section 3, we provide the preliminary theory necessary for our analysis. We formally derive a bound on the change in the allele frequency spectrum due to a change in the population size history in Section 4, both for infinite and finite genome length, and discuss the relationship with Terhorst and Song (2015).

2. The diffusion approximation and the identifiability problem

The evolution of the allele frequency spectrum $P(y,t)$ for a large randomly mating population with Wright–Fisher reproduction can be modelled through a diffusion process (Kimura, 1964), which we write as

$$\frac{\partial}{\partial t}P(y,t)=\frac{1}{2}\frac{\partial^2}{\partial y^2}\left[\frac{y(1-y)}{2N(t)}P(y,t)\right]+2N(t)\mu\,\delta\left(y-\frac{1}{2N(t)}\right) \tag{1}$$

where $0<y<1$ is the allele frequency in the population, μ is the mutation rate per generation, $N(t)$ is the size of the population at time t (measured in generations), and $\delta(\cdot)$ is the Dirac delta function. We assume that the population size is large enough that the frequency y can be approximated by a continuous variable. In this formulation, the first term on the right-hand side describes the effect of genetic drift and the second describes that of new mutations entering the population with initial frequency $1/2N(t)$. The diffusion equation can also be written in "genetic" or "diffusion" time, $\tau(t)=\int_0^t dt'/2N(t')$, in which drift occurs at a constant rate, i.e.,

$$\frac{\partial}{\partial\tau}P(y,\tau)=\frac{1}{2}\frac{\partial^2}{\partial y^2}\left[y(1-y)P(y,\tau)\right]+\left[2\tilde{N}(\tau)\right]^2\mu\,\delta\left(y-\frac{1}{2\tilde{N}(\tau)}\right) \tag{2}$$

where $\tilde{N}(\tau)=N(t(\tau))$.

The frequency spectrum $P(y,\tau)$ depends on $\tilde{N}(\tau)$ and therefore contains information about the population size history. The problem of identifiability of demographic histories, raised by Myers et al. (2008), is that two different histories $\tilde{N}_1(\tau)$ and $\tilde{N}_2(\tau)$ can lead to identical frequency spectra. Concretely, Myers et al. (2008) considered functions $\tilde{N}_2(\tau)$ of the form

$$\tilde{N}_2(\tau)=\tilde{N}_1(\tau)+\alpha N_0\,F(\tau) \tag{3}$$

where α is a constant and N_0 denotes the current population size. They showed that histories $\tilde{N}_1(\tau)$ and $\tilde{N}_2(\tau)$ would lead to identical allele frequency spectra if the function $F(\tau)$ obeys

$$\int_0^\infty F(\tau)\,e^{-\lambda_i\tau}d\tau=0\quad\text{with}\quad\lambda_i=(i+1)(i+2)/2\quad\text{for}\quad i\in\{0,1,2,\ldots\}. \tag{4}$$

They also showed that such functions exist and constructed an example, $F_2(\tau)=\int_0^\tau f_0(\tau-x)f_1(x)dx$ where $f_0(\tau)=\exp(-1/\tau^2)$ and $f_1(\tau)=\left[\cos(\pi^2/\tau)\exp(-\tau/8)\right]/\sqrt{\tau}$, which is displayed in the left panel of Fig. 1 (they set $\alpha=-9$ to ensure that $\tilde{N}_2(\tau)$ is strictly positive). Since $\tilde{N}_2(\tau)$ cannot be distinguished from $\tilde{N}_1(\tau)$ based on the allele frequency spectrum, the inference problem is poorly determined.

However, Bhaskar and Song (2014) pointed out that adding multiples of $F_2(\tau)$ to any smooth history $\tilde{N}_1(\tau)$ leads to unrealistic population histories that oscillate increasingly rapidly as $\tau\to 0^+$. In fact, they showed that any function satisfying Eq. (4) must exhibit an infinite number of sign changes. As such, any population size history function constructed by linear combinations of $F(\tau)$ is biologically unrealistic. They further proved that there is a unique solution to the inference problem in a very general class of realistic model functions. Thus, their argument offers the reassuring message that histories can, in fact, be uniquely reconstructed if we limit ourselves to biologically plausible histories and have sufficient data.

But how much data do we need? The oscillations near $\tau=0$ in $F_2(\tau)$ are so small that they are barely noticeable in the inset within the left panel of Fig. 1 [for example, $F_2(\tau=0.5)\sim 10^{-12}$]. It seems unlikely that these minuscule oscillations are relevant to our ability to reconstruct demographic histories. We expect (and show below) that replacing unrealistic small-amplitude oscillations by a realistic constant value would have an insignificant effect on the resulting spectrum. The resulting history is very different from a constant-sized history but would produce a nearly identical spectrum. If we formulate the inference problem in terms of model likelihoods, the Myers et al. construction shows that the likelihood surface is exactly flat along some directions in the space of all histories. Bhaskar and Song show that such flat directions are not