Theoretical Population Biology xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Theoretical Population Biology

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

Computing the joint distribution of the total tree length across loci in populations with variable size

Alexey Miroshnikov [a,c], Matthias Steinrücken [b,c,*]

[a] *University of California, Los Angeles, Department of Mathematics, United States*
[b] *University of Chicago, Department of Ecology and Evolution, United States*
[c] *University of Massachusetts Amherst, Department of Biostatistics and Epidemiology, United States*

ARTICLE INFO

Article history:
Received 23 November 2016
Available online xxxx

MSC:
92D10
60J27
60J28
35L40

Keywords:
Coalescent theory
Variable population size
Hyperbolic systems of PDEs

ABSTRACT

In recent years, a number of methods have been developed to infer complex demographic histories, especially historical population size changes, from genomic sequence data. Coalescent Hidden Markov Models have proven to be particularly useful for this type of inference. Due to the Markovian structure of these models, an essential building block is the joint distribution of local genealogical trees, or statistics of these genealogies, at two neighboring loci in populations of variable size. Here, we present a novel method to compute the marginal and the joint distribution of the total length of the genealogical trees at two loci separated by at most one recombination event for samples of arbitrary size. To our knowledge, no method to compute these distributions has been presented in the literature to date. We show that they can be obtained from the solution of certain hyperbolic systems of partial differential equations. We present a numerical algorithm, based on the method of characteristics, that can be used to efficiently and accurately solve these systems and compute the marginal and the joint distributions. We demonstrate its utility to study the properties of the joint distribution. Our flexible method can be straightforwardly extended to handle an arbitrary fixed number of recombination events, to include the distributions of other statistics of the genealogies as well, and can also be applied in structured populations.

1. Introduction

Unraveling the complex demographic histories of humans or other species and understanding their effects on contemporary genetic variation is a central goal of population genetics. In addition to advancing our knowledge of the evolutionary processes that shape genomic variation, demographic inference is also an important step towards understanding disease related genetic variation. Recent rapid population growth, for example, severely affects the distribution of rare genetic variants (Keinan and Clark, 2012), which have been linked to complex genetic diseases. Moreover, ancient and contemporary population structure can lead to the accumulation of private genetic variation in certain sub-populations.

Methods to study genetic variation, or perform inference, in populations with varying size or more complex demographic histories have been developed based on the Wright–Fisher diffusion, describing the evolution of population allele frequencies forward in time (Griffiths, 2003; Živković, et al., 2015; Gutenkunst et al., 2009; Excoffier et al., 2013), or the Coalescent process, a model for the genealogical relationship in a sample of individuals (Griffiths and Tavaré, 1994; Griffiths and Marjoram, 1996; Griffiths and Tavaré, 1998; Živković and Wiehe, 2008; Bhaskar et al., 2015; Kamm et al., 2017). A powerful representation of genetic variation data that has been used in this context is the Site-Frequency-Spectrum. In this representation, however, any linkage information present in the genetic data is ignored. With the increasing availability of full-genomic sequence data, linkage information is more readily available, and approaches based on Coalescent Hidden Markov Models (HMM) that use this linkage information have proven to be particularly successful for demographic inference and other population genetic applications.

In a population-sample of genomic sequences, the genealogical relationships vary along the genome, due to intra-chromosomal recombination. The Coalescent-HMMs approximate the intricate correlation structure between these local genealogical trees by a Markov chain, the Sequentially Markovian Coalescent (Wiuf and Hein, 1999; McVean and Cardin, 2005). Due to the Markovian structure of the SMC-approximation, an essential building block is thus the transition or joint distribution of these local genealogies at two neighboring loci. In a sample of size two, the local genealogies are simple trees with two leaves, that is, one-dimensional objects at each locus. The transitions can be readily computed, and Li and

[*] Corresponding author at: University of Chicago, Department of Ecology and Evolution, United States.
E-mail address: steinrue@uchicago.edu (M. Steinrücken).

http://dx.doi.org/10.1016/j.tpb.2017.09.002
0040-5809/© 2017 Elsevier Inc. All rights reserved.

Durbin (2011) employed this framework to develop a powerful approach to infer population size history. Moreover, Dutheil et al. (2009) used Coalescent-HMMs to explore the divergence patterns between humans and great apes, using up to 4 genomic sequences, one for each species. However, due to the increase in complexity of the local genealogies with increasing sample size, these approaches cannot be generalized efficiently to larger sample sizes.

For large sample sizes, approaches that use Monte-Carlo Markov Chain techniques (Rasmussen et al., 2014), suitable composite likelihood frameworks (Sheehan et al., 2013; Steinrücken et al. 2015), or representations of the local genealogical trees by lower-dimensional summaries (Schiffels and Durbin, 2014; Terhorst et al., 2017) have been developed. In the latter, the choice on how to represent the local genealogical trees affects the performance of the inference procedure. Li and Durbin (2011) observed that using the coalescence time between two lineages lacks information in the more recent past, whereas using the first coalescence time in a large sample is less accurate for ancient times (Schiffels and Durbin, 2014). A promising low-dimensional representation is the total branch length of the genealogical tree at each locus. In expectation, this quantity grows without bound as the sample size increases, thus retaining not only information about ancient events, but also about the more recent dynamics. However, to implement a Coalescent-HMM inference framework using the tree length, it is crucial to efficiently compute the joint distribution of the total tree length at two neighboring loci.

Thus, in this paper, we present a novel efficient and accurate method to numerically compute the joint distribution of the total branch length of the genealogical trees at two neighboring loci for a sample of arbitrary size n in populations of varying size, as well as the single-locus marginal distribution. To our knowledge, no method to compute these distributions has been presented in the literature to date that can be applied to arbitrary sample sizes. Moreover, even computing the marginal distribution of the total tree length at a single locus has only received limited attention (Pfaffelhuber et al., 2011; Wiuf and Hein, 1999). We present analytical details and numerical results for the case of at most one recombination event separating the two loci, but our methodology can be readily extended to handle an arbitrary, but fixed, maximal number of recombination events, by suitably augmenting the underlying process.

The inter-coalescent times $T_k^{(n)}$, that is the time period during which k lineages persist in the genealogical tree for a sample of size n can be used to compute the total branch length at a single locus as

$$\mathcal{L} = \sum_{k=2}^{n} kT_k^{(n)}, \tag{1.1}$$

since in the period $T_k^{(n)}$, k lineages contribute towards the total length. In the case of a panmictic population of constant size, formulas for the first two moments of the total tree length can be readily obtained using standard arguments for sums of the independently exponentially distributed random variables $T_k^{(n)}$. Furthermore, $\mathcal{L}$ is distributed like the maximum of $k-1$ exponential variables with intensity $\frac{1}{2}$ (Wiuf and Hein, 1999, p. 255). However, non-constant population size histories introduce intricate dependencies among the inter-coalescent times, and thus it is not straightforward to generalize this approach. Polanski et al. (2003) introduced a method to compute the expected inter-coalescence times under variable population size. However, the coalescence rates of ancestral lineages in the genealogical process depend on past population sizes, whereas the rate for ancestral recombination is constant along each ancestral lineage. The approach of Polanski et al. (2003) depends on the fact that all rates of the process are rescaled uniformly with the same factor, and thus it cannot be extended to the case when ancestral recombination between two linked loci is taken into account.

Ferretti et al. (2013) used another approach to investigate the correlation between the times to the most recent common ancestor at two neighboring loci. The authors approached the problem using coalescent arguments to quantify the changes recombination induces on the local trees, but it is unclear how to generalize their approach efficiently to the total length of the genealogical trees. Furthermore, Li and Durbin (2011) presented analytic formulas for the joint distribution of the local genealogies for a sample of size two under variable population size, but these cannot readily be extended to an arbitrary sample size n. Eriksson et al. (2009) presented similar analytic formulas for a population of constant size and explored more complex demographic scenarios using simulations. Introducing suitable Markov chains, Hobolth and Jensen (2014) investigated the transitional distribution of the local genealogies for samples of size 4, and discussed approximations for larger sample sizes. These Markov chains are closely related to our methodology, but our focus is on exact computations for large sample sizes.

Although we focus on the total tree length under variable population size in a single panmictic population in this paper, our approach can be extended to compute the transition densities for the coalescence time in a sample of size two (Li and Durbin, 2011), the coalescence time of two distinguished lineages (Terhorst et al., 2017), and the time of the first coalescent event amongst the sampled sequences (Schiffels and Durbin, 2014). Furthermore, our method can be generalized to multiple sub-populations related by a complex demographic history (see discussion in Section 5).

This article is structured as follows. In Section 2, we introduce the requisite notation and the stochastic processes that are involved in computing the marginal and joint distributions. We further introduce a hyperbolic system of partial differential equations (PDEs) in Section 3 that can be solved to compute the distributions of interest. We provide a proof of the main proposition used to derive these equations in Appendix A. In Section 3, we also provide the details of our novel numerical algorithm based on the method of characteristics that can be used to efficiently compute the solutions to these PDEs. We demonstrate the accuracy of the method, and study the properties of the joint distribution function in Section 4. Finally, we discuss the future applications and extensions of this method in Section 5.

2. Background and notation

In this section, we will introduce the necessary background and notation for the stochastic processes that we employ to compute the marginal and joint distribution of the length of the genealogical trees. We will also provide some details about computing the distribution of these processes, since our main result extends upon the underlying ideas.

2.1. Ancestral process at a single locus

The genealogical relationship of a sample of n haploid individuals in a panmictic population of constant size is commonly modeled using Kingman's coalescent (Kingman, 1982; Wakeley, 2008), and this process and its extensions have found widespread applications. It is a Markov process that describes the dynamics of the ancestral lineages of the sample backwards in time. Here we focus on the ancestral process $A(t)$ (Tavaré and Zeitouni, 2004, Chapter 4.1). This coarser process records only the number of ancestral lineages in the coalescent process at time t before present, which is sufficient to compute the total branch length of the coalescent tree. The initial number of lineages is equal to the sample size n. Furthermore, at time t, each pair of lineages