Theoretical Population Biology 81 (2012) 144–157

Contents lists available at SciVerse ScienceDirect

Theoretical Population Biology

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

The structure of allelic diversity in the presence of purifying selection

Michael M. Desai [a,b,c,1], Lauren E. Nicolaisen [a,b,c,1], Aleksandra M. Walczak [d], Joshua B. Plotkin [e,*]

[a] Department of Organismic and Evolutionary Biology, Harvard University, United States
[b] Department of Physics, Harvard University, United States
[c] FAS Center for Systems Biology, Harvard University, United States
[d] CNRS-Laboratoire de Physique Théorique de l'École Normale Supérieure, 24 rue Lhomond, 75005 Paris, France
[e] Department of Biology, University of Pennsylvania, United States

ARTICLE INFO

Article history:
Received 13 September 2011
Available online 16 December 2011

Keywords:
Allelic diversity
Purifying selection
Ewens sampling formula
Linkage

ABSTRACT

In the absence of selection, the structure of equilibrium allelic diversity is described by the elegant sampling formula of Ewens. This formula has helped to shape our expectations of empirical patterns of molecular variation. Along with coalescent theory, it provides statistical techniques for rejecting the null model of neutrality. However, we still do not fully understand the statistics of the allelic diversity expected in the presence of natural selection. Earlier work has described the effects of strongly deleterious mutations linked to many neutral sites, and allelic variation in models where offspring fitness is unrelated to parental fitness, but it has proven difficult to understand allelic diversity in the presence of purifying selection at many linked sites. Here, we study the population genetics of infinitely many perfectly linked sites, some neutral and some deleterious. Our approach is based on studying the lineage structure within each class of individuals of similar fitness in the deleterious mutation-selection balance. Consistent with previous observations, we find that for moderate and weak selection pressures, the patterns of allelic diversity cannot be described by a neutral model for any choice of the effective population site. We compute precisely how purifying selection at many linked sites distorts the patterns of allelic diversity, by developing expressions for the likelihood of any configuration of allelic types in a sample analogous to the Ewens sampling formula.

1. Introduction

In any evolving population, new clonal lineages are constantly being created and destroyed. The balance between the creation of lineages by new mutations and their destruction by natural selection and genetic drift determines the statistics of the clonal structure of the population. In the absence of natural selection, Ewens (1972) computed an elegant sampling formula describing the clonal structure of a neutral population, and explained how the allelic (i.e. lineage) configuration in a sample of individuals from the population provides a window into this clonal structure.

Natural selection distorts the clonal structure of a population away from this neutral expectation. Of particular interest is purifying (negative) selection against many linked deleterious mutations ("background selection"). Recent evidence has suggested that this may be generally important in a wide range of populations (see Hahn (2008) for a recent review). In this paper, we explore how this type of selection alters the clonal (i.e. allelic) structure of a population. Our analysis leads to a generalization of the Ewens sampling formula to situations involving background selection.

Over the past few decades, numerous authors have studied allelic diversity in infinite-alleles frameworks that incorporate selection. Li (1977) and Watterson (1978) introduced models in which alleles may have a few different selective effects. Li (1978) and others (Li, 1979; Ewens and Li, 1980; Griffiths, 1983) analyzed the structure of allelic diversity in these models. More recent work has analyzed a very general model of selection introduced by Ethier and Kurtz (1987), which allows for diverse types of selection pressures (Ethier and Kurtz, 1994; Joyce and Tavare, 1995; Grote and Speed, 2002; Joyce, 1995). This work has helped us understand the general effects of selection in distorting the frequency spectrum of sampled alleles. However, the models these authors have analyzed cannot be directly connected to a concrete description of mutations and selection occurring at specific sites. Rather, they assume that each new mutation creates a new allele whose fitness is completely independent of the fitness of its parent. In other words, there is no sense of relatedness among

* Correspondence to: Department of Biology, University of Pennsylvania, 219 Lynch Labs, 433 S. University Avenue, Philadelphia, PA 19104, United States.
E-mail address: jplotkin@sas.upenn.edu (J.B. Plotkin).

[1] These authors contributed equally to this manuscript.

0040-5809/$ – see front matter
doi:10.1016/j.tpb.2011.12.002

alleles, or of a correlation in fitness between closely related alleles. Etheridge and Griffiths (2009) and Etheridge et al. (2010) have more recently derived a coalescent dual of the Moran process with an arbitrary number of types, mutation rates between types, and genic selection coefficients, but it is not clear how this corresponds to selection acting on some fraction of an infinite number of specific sites.

In this paper we take a different approach, based on the specific model of linked sites described by Charlesworth et al. (1993) and Hudson and Kaplan (1994). That is, we imagine that each individual has a genome comprised of many neutral and many negatively selected sites. The fitness of each individual is determined by the number of mutations it carries at the negatively selected sites. We make the infinite-sites assumption that no two mutations at the same site ever segregate simultaneously. This is also an infinite-alleles model, but it is based on a specific model of mutations at individual sites, and the fitness of each new allele depends on the fitness of its parent.

Earlier studies have investigated the effects of purifying selection in models identical or closely related to the one we consider here. Charlesworth et al. (1993) introduced a model essentially identical to the one we analyze here, and Kaplan et al. (1988) and Hudson and Kaplan (1994) developed a simple algorithm which can be used to recursively compute how purifying selection alters the structure of genealogies. Hudson and Kaplan (1995) and Gordo et al. (2002) further developed this idea, resulting in a simple computational method for sampling genealogical relationships in the presence of background selection. Related simulation and analytical work has further characterized the structure of genealogies and the statistics of genetic diversity at the level of individual sites in this or closely related models (McVean and Charlesworth, 2000; Seger et al., 2010; Charlesworth et al., 1993; Comeron and Kreitman, 2002; Comeron et al., 2008; Barton and Etheridge, 2004). However, this earlier work does not provide an analytic description of lineage structure, or sampling formulas for allelic diversity in the presence of purifying selection on many linked sites.

In this paper, we explicitly analyze the lineage structure, and we derive a selected version of the Ewens sampling formula. We begin by noting that the balance between mutations at deleterious sites and selection against them leads to a steady state mutation-selection balance (Haigh, 1978). Our approach is to study the structure of lineages within this steady state, using the Poisson Random Field (PRF) method developed by Sawyer and Hartl (1992). We show that this lineage structure can alternatively be derived using a retrospective approach, by considering the probabilities of mutation and coalescence events in the ancestry of each individual; these probabilities are calculated by Hudson and Kaplan (1994) and Gordo et al. (2002) (and implicitly in a related context by Barton and Etheridge (2004)). Our description of lineage structure is thus precisely consistent with the analysis of genealogical structures in this earlier work. Finally, we use our description of lineage structure to calculate sampling formulas for allelic diversity, and compare our predictions to the results of Monte Carlo simulations.

Provided that selection is strong and deleterious mutation rates are sufficiently small, our results show that the effect of background selection on allelic diversity is to reduce the effective population size without otherwise distorting the lineage structure. Our results are thus consistent with the effective population size approximation to background selection proposed by Charlesworth et al. (1993). For weaker selection, however, or higher mutation rates, the effective population size approximation breaks down, and the effects of background selection become more complex. We show that in this case the allelic diversity cannot be described by neutral theory with some appropriately chosen effective population size. This is consistent with earlier observations that background selection leads to distortions in the structure of genealogies (McVean and Charlesworth, 2000; Seger et al., 2010; O'Fallon et al., 2010; Comeron and Kreitman, 2002; Comeron et al., 2008; Barton and Etheridge, 2004; Gordo et al., 2002; Hermisson et al., 2002; Williamson and Orive, 2002). Our analysis here allows us to compute precisely how these distortions due to purifying selection at many linked sites alter patterns of allelic diversity, and hence provides an analytical framework for exploring where statistical power may lie to distinguish purifying selection from neutrality.

Our approach relies on the assumption that we can describe the distribution of fitnesses within the population with the steady state mutation-selection balance. In particular, we neglect fluctuations within this balance. We note that the PRF and retrospective approaches depend somewhat differently on this key approximation, which offers some insight into the role of fluctuations in our model. We analyze the validity of this approximation in more detail below, and describe a correction for some aspects of the effects of fluctuations in the PRF formalism, which allows us to make a precise correspondence with the retrospective approach. Related to this approximation, we also neglect the effects of Muller's ratchet. We discuss this approximation in detail in the Discussion. We further test the validity of our analysis via Monte Carlo simulations; we find that these approximations are reasonable across a broad parameter regime spanning weak and strong selective pressures.

Our analysis in this paper is limited to allelic diversity, and it does not address the degree of relatedness among sampled alleles. In other words, our analysis only tells us the probability that individuals are genetically identical, not the distribution of the number of specific sites at which individuals may differ. Our results are thus not directly comparable to the work described above, which makes predictions about expected diversity at the level of individual sites. However, while our allele-based results provide an incomplete picture of genetic diversity within the population, they do provide a useful perspective on how purifying selection distorts patterns of molecular evolution. Most importantly, we are able to make precise analytical predictions about how purifying selection distorts allelic diversity, in ways that cannot be described by a single reduced effective population size.

2. Model

We imagine a finite haploid population of constant size N. Each haploid genome has a large number of sites, which begin in some ancestral state and mutate at a constant rate. Each mutation is either neutral or confers some fitness disadvantage s (where by convention $s > 0$). We assume an infinite-sites framework, so there is negligible probability that two mutations segregate simultaneously at the same site.

We assume that there is no epistasis for fitness, and that each deleterious mutation carries fitness cost s, so that the fitness of an individual with k deleterious mutations is $w_k = (1-s)^k$. Since we assume that $s \ll 1$, we will often approximate w_k by $1 - sk$. Later we comment briefly on extensions to our method to consider the case when the selection coefficient of a deleterious mutations is drawn from some fixed distribution.

The population dynamics are assumed to follow the diffusion limit of the standard Wright–Fisher model. That is, we assume that deleterious mutations occur at a genome-wide rate U_d per individual per generation (with deleterious mutations assumed to be decoupled from selection). We define $\theta_d/2 \equiv NU_d$, the per-genome scaled deleterious mutation rate. Similarly, neutral mutations occur at a rate U_n per individual per generation, and we analogously define $\theta_n/2 \equiv NU_n$. We assume that each newly