Theoretical Population Biology 106 (2015) 71–82

Contents lists available at ScienceDirect

Theoretical Population Biology

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

Inference of directional selection and mutation parameters assuming equilibrium

Claus Vogl [a,*], Juraj Bergman [b,c]

[a] Institute of Animal Breeding and Genetics, Veterinärmedizinische Universität Wien, Veterinärplatz 1, A-1210 Vienna, Austria
[b] Institute of Population Genetics, Veterinärmedizinische Universität Wien, Veterinärplatz 1, A-1210 Vienna, Austria
[c] Vienna Graduate School of Population Genetics, Veterinärmedizinische Universität Wien, Veterinärplatz 1, A-1210 Vienna, Austria

ARTICLE INFO

Article history:
Received 10 July 2015
Available online 24 October 2015

Keywords:
Biallelic mutation–selection–drift model
Small scaled mutation rate
Equilibrium density
Inference

ABSTRACT

In a classical study, Wright (1931) proposed a model for the evolution of a biallelic locus under the influence of mutation, directional selection and drift. He derived the equilibrium distribution of the allelic proportion conditional on the scaled mutation rate, the mutation bias and the scaled strength of directional selection. The equilibrium distribution can be used for inference of these parameters with genome-wide datasets of "site frequency spectra" (SFS). Assuming that the scaled mutation rate is low, Wright's model can be approximated by a boundary-mutation model, where mutations are introduced into the population exclusively from sites fixed for the preferred or unpreferred allelic states. With the boundary-mutation model, inference can be partitioned: (i) the shape of the SFS distribution within the polymorphic region is determined by random drift and directional selection, but not by the mutation parameters, such that inference of the selection parameter relies exclusively on the polymorphic sites in the SFS; (ii) the mutation parameters can be inferred from the amount of polymorphic and monomorphic preferred and unpreferred alleles, conditional on the selection parameter. Herein, we derive maximum likelihood estimators for the mutation and selection parameters in equilibrium and apply the method to simulated SFS data as well as empirical data from a Madagascar population of *Drosophila simulans*.

1. Introduction

Mutation creates the genome-wide sequence variation upon which other population genetic forces act. DNA sequence data, in principle, offer the opportunity to infer important population genetic parameters—effective population sizes, mutation rates, and selection coefficients. Together with weak directional selection, mutation bias is thought to affect the nucleotide base composition, such that different genomic regions may vary within a species, *e.g.* between classes of sites such as short introns and fourfold degenerate sites, and among closely related species (*e.g.* Singh et al., 2005, 2007, 2009; Parsch et al., 2010; Clemente and Vogl, 2012b,a). The mutation parameters are most easily inferred from data of putatively neutral or nearly-neutral sequences. The nucleotide base composition is also influenced by other population genetic forces, *e.g.* the recombination rate, and features, *e.g.* X- or autosomal linkage. Thus improved methods of inference of mutation and selection parameters may also help understand these forces and features.

Typically, inference methods rely on the scaled mutation rates being so small that mutations can be assumed to be irreversible and always occur at new sites; usually these methods also require inference of ancestral states (but see RoyChoudhury and Wakeley, 2010; Vogl and Clemente, 2012). Here, we relax this assumption somewhat using a reversible mutation model. To do this, we rely on theory developed by Wright (1931). His formula for mutation–selection–drift equilibrium

$$\Pr(x|\alpha,\theta,\gamma)=\frac{e^{\gamma x}x^{\alpha\theta-1}(1-x)^{(1-\alpha)\theta-1}}{\int_0^1 e^{\gamma x}x^{\alpha\theta-1}(1-x)^{(1-\alpha)\theta-1}\,dx} \tag{1}$$

has stood the test of time (here x is the population proportion of the preferred allele, γ the scaled selection coefficient, α the mutation bias, and $\theta=\mu N$ the scaled mutation rate; in the following, we will often set $\beta=1-\alpha$). If $\gamma \ll 1$ mutation dominates (the neutral

[*] Corresponding author.
E-mail addresses: claus.vogl@vetmeduni.ac.at (C. Vogl), juraj.bergman@vetmeduni.ac.at (J. Bergman).

http://dx.doi.org/10.1016/j.tpb.2015.10.003
0040-5809/© 2015 Elsevier Inc. All rights reserved.

region), such that a beta describes the equilibrium distribution

$$\Pr(x|\alpha,\theta) = \frac{\Gamma(\theta)}{\Gamma(\alpha\theta)\Gamma(\beta\theta)}\, x^{\alpha\theta-1}(1-x)^{\beta\theta-1}. \tag{2}$$

If $\gamma > 4$ selection dominates and nearly no unpreferred alleles are found in the population (Ohta and Gillespie, 1996). The intermediate case, where $\gamma \approx 1$, corresponds to the "nearly-neutral region" (Ohta, 1979; Ohta and Gillespie, 1996). In this parameter range, directional selection opposing mutation bias may increase equilibrium variability (McVean and Charlesworth, 1999).

Paradigmatic datasets for application of the biallelic mutation–selection–drift model come from "site frequency spectra" (SFS) of short intron sequences and fourfold degenerate sites in *Drosophila* species of the *melanogaster* subgroup. Sites between positions 8 and 30 in short introns are assumed to have no function except as spacers and are thus considered to evolve neutrally, or very nearly so (Parsch et al., 2010). For the theory to be applicable to *Drosophila* data, A and T nucleotides can be grouped together and contrasted with C and G nucleotides. This binary classification is justified by the assumption that mutation is not strand-specific, but is often either AT or CG biased, while selection or biased gene conversion may favor CG over AT. For a sample of M haplotypes and a specific site class, data can be represented as a site frequency spectrum (SFS) by defining L_y to be the number of genomic sites that have y CG bases and $(M-y)$ AT bases. Obviously, $0 \le y \le M$ and the total number of sites is $L = \sum_{y=0}^{M} L_y$.

Neither the general formula (1) nor the beta distribution (2) have often been used for inference of population genetic parameters, in spite of being known for more than 80 years and their importance in population genetics. Rather, inference has generally been based on approaches that explicitly or implicitly assume small scaled mutation rates θ. With the infinite sites model, infinitely many sites may be hit by mutation at a finite rate, such that each site is hit only once (Kimura, 1964; Kimura and Ohta, 1969; Watterson, 1975). Furthermore, it is usually assumed that the ancestral state is known via outgroup information, *i.e.* alleles can be polarized into ancestral and derived, which is only possible if mutations are rare enough to make multiple hits of the same site unlikely. The well-known Ewens–Watterson estimator of scaled mutation rate (Ewens, 1974; Watterson, 1975), $\hat{\theta}_w = L_p/(L\sum_{y=1}^{M-1} 1/y)$, where the number of polymorphic sites $L_p = \sum_{y=1}^{M-1} L_y$, is based on the infinite sites model. The Ewens–Watterson estimator is generally unbiased; if sites are unlinked, it is also the maximum likelihood estimator of θ. If assumptions are met, $\hat{\theta}_w$ corresponds to the "expected heterozygosity", *i.e.* the proportion of polymorphism in a sample of size $M = 2$.

Similar to the infinite sites model, applications of the Poisson Random Field (PRF) model to population genetics explicitly or implicitly assume small scaled mutation rates. Often, theory is based on irreversible mutation models and applied to sequence variation, which requires knowledge of ancestral states and an unlimited and unvarying supply of sites (*e.g.* Sawyer and Hartl, 1992; Bustamante et al., 2001, 2003; Williamson et al., 2004). Some of these models allow for a distribution of selection coefficients or arbitrary dominance (Bustamante et al., 2003; Williamson et al., 2004). As far as we are aware, among the theory based on the PRF model only RoyChoudhury and Wakeley (2010) do not assume outgroup information, but rather start from a Taylor series expansion in θ of distributions (1) and (2). Nevertheless, the estimator of the scaled mutation rate that RoyChoudhury and Wakeley (2010) derive is essentially identical to the Ewens–Watterson estimator. Starting from a Moran model, Vogl and Clemente (2012) derive a similar equation also for the case with directional selection.

In *Drosophila melanogaster* and *Drosophila simulans*, the ratio (AT):(CG) in short introns is about 2:1. In the same species, this ratio reverses to about 1:2 in fourfold degenerate sites. Since the mutation process is unlikely to differ between the two site classes, directional selection favoring CG nucleotides in fourfold degenerate sites is probably the force behind this observation (Hershberg and Petrov, 2008; Clemente and Vogl, 2012b,a). With low effective mutation rates, such that only a single mutation segregates, the ratio of the unpreferred to the preferred allele is about $\beta : \alpha e^{\gamma}$; a result that can already be derived from Wright (1931). Assuming no directional selection in short introns, the scaled selection coefficient favoring CG over AT in fourfold degenerate sites is therefore about $\gamma \approx \log(4) = 1.39$. Obviously γ is of the order one in this case, *i.e.* in the nearly-neutral region. Furthermore, the strength of the directional selective force and of the mutation bias are balanced in a way that short intron and fourfold degenerate sites are about equally polymorphic (Parsch et al., 2010; Clemente and Vogl, 2012b,a). Even more importantly, the amount of polymorphism is so low ($\hat{\theta} \approx 0.02$) that the assumption of low scaled θ is met, according to simulations in Vogl and Clemente (2012).

Herein, we will derive the sampling distribution (likelihood) of a biallelic locus in mutation–selection–drift equilibrium in the nearly-neutral range assuming low scaled mutation rates θ. For most organisms and datasets (*e.g. Drosophila* species and mammals), this restriction does not seem to be a problem. Furthermore, the assumption of low scaled mutation rates allows us to draw parallels to earlier methods of inference of population genetic parameters, which are based on models that explicitly or implicitly assume low scaled mutation rates θ. To derive this likelihood, we will put forward a boundary-mutation model. For $\gamma = 0$ the sampling distribution of the boundary-mutation model is identical to that of a Taylor series of the general model up to first order in θ. For $\gamma \neq 0$, it provides computationally feasible maximum likelihood estimators of all three parameters, α, θ, and γ. A set of functions used for parameter inference was implemented in the "R"-software (R Core Team, 2014) and made available for further usage. We apply the inference method to simulated datasets and to data from short introns and fourfold degenerate sites in a population sample of *Drosophila simulans* from Madagascar (Rogers et al., 2014).

1.1. Assumptions

The following is assumed throughout:

Assumption 1. Allele frequency data of biallelic loci (sites), *i.e.* SFS data, are available from $L < \infty$ loci, indexed by $1 \le l \le L$.

Assumption 2. The allelic proportions x_l of the preferred allele at each of the L sites are independently and identically drawn from a mutation–selection–drift equilibrium solution.

Assumption 3. The likelihood of the allelic frequencies y_l in the sample is binomial with parameters x_l and identical sample size M.

Note that dropping the assumption of constant M for all L only increases the complexity of notation.

2. Inference with a general mutation–selection–drift model

This section is mainly a review of inference with a general biallelic mutation–selection–drift model in equilibrium. So far,