Theoretical Population Biology 78 (2010) 118–122

Contents lists available at ScienceDirect

Theoretical Population Biology

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

Sufficiency of the number of segregating sites in the limit under finite-sites mutation

Arindam RoyChoudhury [a,*], John Wakeley [b]

[a] Department of Biostatistics, Columbia University, New York, NY 10032, United States

[b] Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, United States

ARTICLE INFO

Article history:
Received 16 January 2010
Available online 2 June 2010

Keywords:
Ewens' sampling formula
Infinitely-many-alleles model
Infinitely-many-sites model
Maximum likelihood estimator
Mutation parameter
Number of segregating sites
Poisson Random Field
Sufficient statistic
Watterson's estimator

ABSTRACT

We show that the number of segregating sites is a sufficient statistic for the scaled mutation parameter (θ) in the limit as the number of sites tends to infinity and there is free recombination between sites. We assume that the mutation parameter at each site tends to zero such than the total mutation parameter (θ) is constant in the limit. Our results show that Watterson's estimator is the maximum likelihood estimator in this case, but that it estimates a composite parameter which is different for different mutation models. Some of our results hold when recombination is limited, because Watterson's estimator is an unbiased, method-of-moments estimator regardless of the recombination rate. The quantity it estimates depends on the details of how mutations occur at each site.

1. Introduction

One of the main goals of population genetics is to estimate parameters such as the population scaled mutation rate. For a diploid organism, this parameter is defined as $\theta = 4N\mu$, where N is the population size and μ is the mutation rate. For a haploid organism, it is defined as $2N\mu$. The mutation rate μ is defined in one of two different ways: as a rate per genetic locus or as a rate per nucleotide site. With the additional concept of effective population size – whereby N is replaced by N_e – the parameter $\theta = 4N_e\mu$ has been shown to accurately capture the balance between the introduction of genetic variation by mutation and its loss by random genetic drift in a single large population, nearly irrespectively of the demographic details of the population (Ewens, 1982; Sjödin et al., 2005). In more complicated scenarios, such as when multiple populations are connected by migration, other parameters arise and it is of interest to estimate these as well.

Population parameters are estimated from genetic data that is sampled from the population. Decades ago, genotypes could only be measured indirectly, using methods like protein electrophoresis or restriction-enzyme digests of DNA, while today it is possible to obtain genotypes directly, for example by sequencing DNA. Statistical models for estimating population parameters have been developed for the whole range of types of data, and this accounts for the different definitions of the mutation rate, μ, that appear in the literature. Earlier models are still in use, although most current applications use DNA data and the associated statistical models.

The problem of estimating population parameters from genetic data is complicated by the fact that individuals sampled from the population are related through common ancestors. These ancestries cannot be observed directly, and must be treated as missing data. One result of common ancestry is that genotypes at two or more sites among a sample of individuals may not be independent if the sites are physically close together in the genome. Methods of estimating population parameters vary from simple method-of-moments estimators, based on closed-form analytic expressions, to maximum-likelihood or Bayesian estimators, which "integrate" over ancestries using Monte Carlo methods. Due to the complicated structure of genetic data, induced by common ancestry, the statistical properties of most estimators are known only through simulations.

Here, we focus on one particular statistical property called sufficiency. A sufficient statistic for a parameter is one which captures all of the information in a data set relevant to the estimation of that parameter; specifically such that the likelihood of the parameter does not depend on any other aspect of the data. Although this is but one of several possible measures of the quality of an estimator, studies of sufficiency may be of particular importance

* Corresponding author.
E-mail address: ar2946@columbia.edu (A. RoyChoudhury).

0040-5809/$ – see front matter
doi:10.1016/j.tpb.2010.05.003

in population genetics due to the growing popularity of approximate Bayesian computing (Beaumont et al., 2002) and other methods of inference based on summary statistics rather than the full data. If we want to estimate a given parameter, then knowledge about which aspects of the data contain information about the parameter will facilitate the choice of summary statistics.

Our concern here is much simpler than this. We focus on the standard model of population genetics: the neutral, diploid, monoecious Wright–Fisher model (Fisher, 1930; Wright, 1931). The population scaled mutation rate for this model is $\theta = 4N\mu$ (because $N_e = N$ in this case). Note that, implicitly in what follows, μ is defined as the mutation rate for all the sites that are genotyped in the sample. Following a suggestion made by Ewens (1974), our goal is to show that the number of segregating (or polymorphic) sites in a sample is a sufficient statistic for θ under a version of Kimura's (Kimura, 1969) infinitely-many-sites model.

Two different infinitely-many-sites models have been proposed, one by Kimura (1969) and another by Watterson (1975). In both models, each mutation occurs at a previously unmutated site. Conceptually, multiple mutations at single sites will never occur if θ is finite and there are infinitely-many sites. In contrast, the two models make very different assumptions about recombination, which is the process that mediates the dependence between sites due to common ancestry (Hudson, 1983; Kaplan and Hudson, 1985). Watterson's model assumes that recombination does not occur between sites, while Kimura's model assumes that recombination is so frequent between sites that their genotypes are independent.

In Watterson's infinitely-many-sites model, each new mutation creates a new allele (or haplotype) which is then faithfully transmitted due to the absence of recombination. Thus, if only allelic states are recorded, Watterson's infinitely-many-sites model is equivalent to the infinitely-many-alleles model (Malécot, 1946; Kimura and Crow, 1964). This is not true for Kimura's infinitely-many-sites model because recombination may also create new alleles. Ewens (1972) showed that the number of distinct alleles, k, in the sample is a sufficient statistic for θ under the infinitely-many-alleles model. The sample frequencies of alleles contain no additional information about θ. Note that, if the number of possible alleles is finite and equal to K, then the observed number of alleles k is not a sufficient statistic for θ. Ewens' (Ewens, 1972) result is then seen as a limiting result, as $K \to \infty$.

If, rather than the number of alleles, k, one records the number of segregating sites, k^*, then this too may be used to estimate θ. As Ewens (1974) showed for the infinitely-many-sites model with free recombination and Watterson (1975) showed for the infinitely-many-sites model with no recombination, the quantity

$$\frac{k^*}{\sum_{j=1}^{n-1} 1/j}, \tag{1}$$

where n is the sample size, provides an unbiased estimate of θ. It is straightforward to show that this remains true with any rates of recombination between sites. We will follow the common practice of calling Expression (1) "Watterson's" estimator. The notation k^* is from Ewens (1974), and is equivalent to the quantity S which appears in the literature.

In general, k^* is not a sufficient statistic for θ (Ewens, 1974; Watterson, 1975), but Ewens (1974) suggested that this would be true under Kimura's infinitely-many-sites model. Here, we adopt Kimura's assumption that the sites that are genotyped in the sample are statistically independent of one another, so that the probability of the full data set is equal to the product of the probabilities of the data at each site. Unlike Kimura (1969) (and also Watterson (1975)) who assumed implicitly that the number of sites is infinite, we consider a collection of L sites and obtain the infinitely-many-sites model in the limit $L \to \infty$. We will assume that θ, which again applies to the entire collection of L sites, is finite in the limit, so that the mutation rate at each site tends to zero. Then, each segregating site will be the result of its own unique mutation as in Kimura's model.

For a number of different pre-limiting models, which differ in the mutation process at each site, we find that k^* is a sufficient statistic for θ in the limit $L \to \infty$. This is not true of k^* in the pre-limiting, finitely-many-sites models.

We note that the sufficiency of k^* for θ in the limiting model is implicit in the Poisson Random Field (PRF) models of Sawyer and Hartl (1992). This holds in other versions of PRF models as well. The other versions include Williamson et al. (2004) where dominance is modeled and incorporated, Wakeley (2003) where population subdivision is modeled, and Zhu and Bustamante (2005) where linkage between sites is incorporated; see also Bustamante et al. (2002, 2005) and Sawyer et al. (2003). However, in these models it is assumed at the outset that each mutation occurs at a previously unmutated site, while in our models this is a result that occurs in the limit $L \to \infty$. Desai and Plotkin (2008) recently studied a finitely-many-sites model with selection and symmetric mutation, but did not consider the limit $L \to \infty$.

Our results have implications for the analysis of DNA sequences or other genetic data. In particular, for all the pre-limiting models we consider, we show that the unbiased maximum likelihood estimator (MLE) of θ, based on k^* in the limit $L \to \infty$, is equal to Watterson's estimator Expression (1) *times by a constant* which depends on the pattern of mutation among alleles at each site. If there is more than one mutation parameter involved, then the MLE is a combination of Expressions (1) applied to those parameters (Section 2.3). Therefore, Expression (1) by itself applied to data does not estimate θ, but rather θ times this constant. For example, consider the simple case where $K = 2$ alleles ('1' and '2') are possible at each site but mutation rates may be asymmetric, then Watterson's estimator estimates the harmonic mean of the two mutation rates, θ_{12} and θ_{21}. Since Watterson's estimator is a method-of-moments estimator based on the expected number of segregating sites, this result hold for any levels of recombination between sites. By extension, the "θ" estimated from a sample of DNA sequences will depend on the frequencies of the four nucleotides and the rates of mutation among them, in a way that is not recognized in simple infinitely-many-sites models.

2. Models and theory

We begin this section with a general statement of the model and our main result, Theorem 1. Subsequently, we apply the result to some well known models for the mutation process at each site. We use the word "site" to emphasize the connection to DNA data, but we do not restrict ourselves to DNA-based ($K = 4$) models. For example, in Section 2.1 we allow infinitely-many-alleles mutation at each site. Although the word "locus" might be better in this case, we use site in all cases for simplicity and to underscore one important aspect of our results, which is that all of the models converge to a version of Kimura's infinitely-many-sites model in the limit as the number of sites tends to infinity. We end this section with a brief discussion of non-independent sites.

Consider L sites. In general, we may think of a sample of size n_i taken at site i. Let k_i be the number of allelic types that are observed in the sample and let θ_i be the mutation parameter at the site i. The infinitely-many-sites model with free recombination between sites (hereafter ISM), as it is usually applied to a sample, involves four assumptions:

Assumption 1. The allele frequencies in the L sites are independent of each other.

Assumption 2. The sample size n_i is the same at every site: $n_i = n$ for all $i = 1, 2, \ldots, L$.

Assumption 3. The total mutation parameter θ is finite, i.e. $\sum_{i=1}^{L} \theta_i = \theta < \infty$, even in the limit $L \to \infty$.

Assumption 4. The per-site mutation parameter is the same at every site: $\theta_i = \theta_L = \theta/L$ for all i.

Let us make all four assumptions to study the limiting behavior of the joint distribution of the allele-counts at the sites. Some of these assumptions may be relaxed, as noted later in this section. For the sake of clarity and simplicity, however, we will make all four assumptions here. Again, the fundamental idea is to consider a large number of such sites, all independent of one another, and for which θ_L is small.

Now, let k^* be the number of segregating sites, which is the number of sites among our L sites that have at least two alleles in the sample. Let $k_0^* = L - k^*$ be the number of monomorphic sites and $k_j^*, j = 1, 2, \ldots, [n/2]$, be the number of sites with exactly two-alleles with counts j and $n-j$ ($j,\ n-j > 0$). Note that k_j^* includes both sites at which the mutant base is in count j (and the ancestral base is in count $n-j$) and sites at which the mutant base is in count $n-j$. Thus, we assume that the ancestral state at each site is unknown, which is appropriate because this information is essentially never available. Let $f_{2,j}(\theta_L)$ be the probability of a site having exactly two alleles with counts j and $n-j$ ($j,\ n-j > 0$) and $f_{>2}(\theta_L)$ be that of a site having more than two alleles.

Theorem 1. *Suppose that we have L independent sites, where each site has the same allele distribution satisfying the following conditions:*

$$f_{2,j}(\theta_L) = \theta_L c_j + O(\theta_L^2) \quad j = 1, 2, \ldots, [n/2], \tag{2}$$

$$f_{>2}(\theta_L) = O(\theta_L^2). \tag{3}$$

Then, in the limit $L \to \infty$ and holding $\theta = L\theta_L$ constant:

(i) *the joint probability distribution of $(k_0^*, k_1^*, \ldots, k_{[n/2]}^*)$ converges to the distribution of a PRF model, (Sawyer and Hartl, 1992) as $L \to \infty$.*

(ii) *the asymptotic distribution of the total number of segregating sites k^* is Poisson with mean $\theta \sum_{j=1}^{[n/2]} c_j$.*

Moreover, if the c_j are known quantities, then in the limit:

(iii) *k^* is sufficient for the parameter θ.*

(iv) *the MLE of θ based on the distribution of k^* is*

$$\hat{\theta} = \frac{k^*}{\sum_{j=1}^{[n/2]} c_j},$$

which is a version of Watterson's estimator.

(v) *each segregating site has exactly two alleles.*

Proof. It follows from Eqs. (2) and (3) that the probability of observing a single allele is

$$f_1(\theta_L) = 1 - \theta_L \sum_{j=1}^{[n/2]} c_j + O(\theta_L^2).$$

Thus, for a given L, the distribution of $k_0^*, k_1^*, \ldots, k_{[n/2]}^*$ differs by $O(L\theta_L^2)$ from the multinomial distribution

$$\Pr(k_0^*, k_1^*, \ldots, k_{[n/2]}^* | L, \theta_L, n) = \binom{L}{k_0^* k_1^* \cdots k_{[n/2]}^*} (1 - c_0\theta_L)^{k_0^*} (c_1\theta_L)^{k_1^*} \cdots (c_{[n/2]}\theta_L)^{k_{[n/2]}^*}, \tag{4}$$

where $c_0 = \sum_{j=1}^{[n/2]} c_j$. As L tends to infinity, with $\theta = L\theta_L$ fixed, $(k_1^*, \ldots, k_{[n/2]}^*)$ converges in distribution to mutually-independent Poissons with parameters $\theta c_j, j = 1, 2, \ldots, [n/2]$ (see, for example, Feller (1970) page 172). Thus, the joint probability distribution of $(k_0^*, k_1^*, \ldots, k_{[n/2]}^*)$ converges to the distribution of a PRF. This proves (i).

The limiting joint probability mass function of $k_1^*, \ldots, k_{[n/2]}^*$ is,

$$\Pr(k_1^*, \ldots, k_{[n/2]}^* | \theta, n) \prod_{j=1}^{[n/2]} e^{-\theta c_j} \frac{(\theta c_j)^{k_j^*}}{k_j^*!} = e^{\left(-\theta \sum_{j=1}^{[n/2]} c_j\right)} \left(\prod_{j=1}^{[n/2]} \frac{c_j^{k_j^*}}{k_j^*!} \right) \theta^{k^*}.$$

Using factorization criterion for sufficiency, k^* is a sufficient statistic for θ in the limiting distribution. This proves (iii).

As $k_1^*, \ldots, k_{[n/2]}^*$ are asymptotically independent Poisson random variables with means $\theta c_1, \theta c_2, \ldots, \theta c_{[n/2]}$ respectively,

$$k^* = \sum_{j=1}^{[n/2]} k_j^*$$

is asymptotically Poisson distributed with mean $\theta \sum_{j=1}^{[n/2]} c_j$. This proves (ii).

It follows that the MLE of θ based on the asymptotic distribution of k^* is

$$\frac{k^*}{\sum_{j=1}^{[n/2]} c_j}.$$

This proves (iv).

The statement (v) then follows from Eq. (3). □

It is easy to show that a version Theorem 1 will still hold if we relax Assumptions 2 and 4 in the following way:

Assumption 2′. Suppose that the sample size is $n_{i'}$ for a fraction $\gamma_{i'}$ of the sites ($i' = 1, 2, \ldots, L'$), where $n_{i'}$ and $\gamma_{i'}$ are known.

Assumption 4′. Suppose that $\theta_{i''} = y_{i''}\theta$ for a fraction $\beta_{i''}$ of all the sites ($i'' = 1, 2, \ldots, L''$), where $y_{i''}$ and $\beta_{i''}$ are known. (The sum of all the site-specific mutation parameter is θ; that is $\sum_{i''} \beta_{i''}\theta_{i''} = \theta$.)

2.1. Infinitely-many alleles

Ewens (1972) discovered a now well known sampling formula under the infinitely-many-alleles model of mutation. In particular,

$$\Pr(k, a_1, a_2, \ldots, a_n | \theta_L, n) = \frac{\theta_L^k n!}{(\theta_L)_{(n)} \prod_{j=1}^{n} j^{a_j} a_j!} \tag{5}$$

is the probability that a sample of size n contains k different allelic types and that a_j alleles are represented j times in the sample, for $j = 1, 2, \ldots, n$, and where

$$(\theta_L)_{(n)} = \theta_L(\theta_L + 1) \cdots (\theta_L + n - 1).$$

Using Eq. (5), we have

$$f_{2,j}(\theta_L) = \theta_L \left(\frac{1}{j} + \frac{1}{n-j} \right) + O(\theta_L^2) \quad \text{if } j \neq n/2, \tag{6}$$

$$f_{2,n/2}(\theta_L) = 2\theta_L/n + O(\theta_L^2), \tag{7}$$

which agrees with the result obtained by Tajima (1989) and Fu (1997) for Watterson's infinitely-many-sites model.

Thus, the infinitely-many-alleles model conforms to the conditions in Eqs. (2) and (3) with $c_j = 1/j + 1/(n-j)$, for $j \neq n/2$, and $c_{n/2} = 2/n$. Therefore, we may apply Theorem 1. Thus, k^* is sufficient for θ. Moreover,

$$\hat{\theta} = \frac{k^*}{\sum_{j=1}^{n-1} 1/j},$$

which is identical to Watterson's estimator, is the MLE of θ based on the limiting distribution. It also follows from Theorem 1 that each segregating site has exactly two alleles in the limit. We do not distinguish between the ancestral and the derived type in Eqs. (6) and (7), but this is not relevant to the inference of the mutation parameters.

2.2. Parent-independent mutation

Next we will consider a multiallelic model with $K(<\infty)$ alleles and 'parent-independent' mutation. Under parent-independent mutation (PIM), the mutation rate from allele i' to allele i does not depend on i', but may depend on i. Let π_i be the probability that the mutated allele is of the type i, given that a mutation has taken place, where $\sum_{i=1}^{K}\pi_i = 1$. Then at each site, the mutation parameter associated with a mutation to allele i is $\theta_L\pi_i$. Note that in this model an allele can "mutate" to an allele of the same type.

Under PIM, the stationary distribution of allele frequencies in the population is Dirichlet with K parameters, and with parameter i equal to $\theta_L\pi_i$ (Wright, 1949). The probability of observing j_i copies of allele i ($i = 1, 2, \ldots, K$) in a sample of size n is multinomial-Dirichlet, with the probability function

$$\begin{aligned} &f(j_1, j_2, \ldots, j_{K-1}; n, K, \theta_L, \pi_1, \pi_2, \ldots, \pi_K) \\ &\quad = \frac{n!}{j_1!j_2!\ldots j_K!}\frac{D_K(j_1+\theta_L\pi_1, j_2+\theta_L\pi_2, \ldots, j_K+\theta_L\pi_K)}{D_K(\theta_L\pi_1, \theta_L\pi_2, \ldots, \theta_L\pi_K)} \\ &\quad = \frac{n!}{j_1!j_2!\ldots j_K!}\frac{\Gamma(j_1+\theta_L\pi_1)}{\Gamma(\theta_L\pi_1)}\frac{\Gamma(j_2+\theta_L\pi_2)}{\Gamma(\theta_L\pi_2)}\cdots \\ &\quad \times \frac{\Gamma(j_K+\theta_L\pi_K)}{\Gamma(\theta_L\pi_K)}\frac{\Gamma(\theta_L)}{\Gamma(n+\theta_L)}. \end{aligned}$$

Note that

$$\frac{\Gamma(j+\theta_L)}{\Gamma(\theta_L)} = (j-1)!\theta_L + O\left(\theta_L^2\right) \quad \text{for } j > 0, \quad \text{and} \tag{8}$$

$$\frac{\Gamma(j_i+\theta_L\pi_i)}{\Gamma(\theta_L\pi_i)} = \pi_i\theta_L(j_i-1)! + O\left(\theta_L^2\right)$$

$$\text{if and only if} \quad j_i > 0. \tag{9}$$

Therefore, the probability of observing j copies of allele i_1 and $n-j$ copies of allele i_2 is

$$\begin{aligned} &\frac{n!}{j!(n-j)!}\frac{\Gamma(j+\theta_L\pi_{i_1})}{\Gamma(\theta_L\pi_{i_1})}\frac{\Gamma(n-j+\theta_L\pi_{i_2})}{\Gamma(\theta_L\pi_{i_2})}\frac{\Gamma(\theta_L)}{\Gamma(n+\theta_L)} \\ &\quad = \frac{n!}{j!(n-j)!}\left((j-1)!\theta_L\pi_{i_1}\right)\left((n-j-1)!\theta_L\pi_{i_2}\right)\left((n-1)!\theta_L\right)^{-1} \\ &\quad = \theta_L\pi_{i_1}\pi_{i_2}\left(\frac{1}{j}+\frac{1}{n-j}\right) + O\left(\theta_L^2\right). \end{aligned} \tag{10}$$

Thus, the probability of observing exactly two alleles of counts j and $n-j$ is

$$\theta_L\left(\sum_{i=1}^{K}\pi_i(1-\pi_i)\right)\left(\frac{1}{j}+\frac{1}{n-j}\right) + O\left(\theta_L^2\right) \quad \text{if } j \neq [n/2] \tag{11}$$

$$\theta_L\left(\sum_{i=1}^{K}\pi_i(1-\pi_i)\right)\frac{2}{n} + O\left(\theta_L^2\right) \quad \text{if } j = [n/2] \tag{12}$$

which has the form of Eq. (2). Also, from Eqs. (8) and (9) the probability of observing at least three alleles in the sample is $O(\theta_L^2)$.

Therefore, we may apply Theorem 1. Each segregating site has exactly two alleles in the limit,

$$c_j = \left(1-\sum_{i=1}^{K}\pi_i^2\right)\left(\frac{1}{j}+\frac{1}{n-j}\right) \quad \text{if } j \neq n/2, \tag{13}$$

$$c_{n/2} = \left(1-\sum_{i=1}^{K}\pi_i^2\right)\frac{2}{n}, \tag{14}$$

and the statistic k^* is Poisson with mean

$$\theta\left(1-\sum_{i=1}^{K}\pi_i^2\right)\sum_{j=1}^{n-1}1/j.$$

If $\pi_1, \pi_2, \ldots, \pi_K$ are known, then k^* is sufficient for θ and the MLE of θ based on the limiting distribution of k^* is

$$\hat{\theta} = \frac{k^*}{\left(1-\sum_{i=1}^{K}\pi_i^2\right)\sum_{j=1}^{n-1}1/j}. \tag{15}$$

Note that Eq. (15) is a multiple of Watterson's estimator. Applied as is, Watterson's estimator Expression (1) would give an estimate of $\theta(1-\sum_{i=1}^{K}\pi_i^2)$, which we may think of as the net rate of mutation to different alleles. This is desirable since it is only for mathematical convenience that the PIM model includes false mutation events between alleles of the same type.

In the case of symmetric mutation, where $\pi_i = 1/K$ for all i, Eq. (15) becomes

$$\hat{\theta} = \frac{k^*}{(1-1/K)\sum_{j=1}^{n-1}1/j},$$

and direct application Expression (1) provides an estimate of $\theta(1-1/K)$, which in this case is exactly the rate of mutation to different alleles. If $K = 4$, then this model is equivalent to the Jukes–Cantor substitution model (Jukes and Cantor, 1969).

Finally, the c_j's in Eqs. (13) and (14) converge to those from the infinitely-many-alleles model as long as

$$\lim_{K\to\infty}\sum_{i=1}^{K}\pi_i^2 = 0. \tag{16}$$

This increases the generality of the results.

2.3. Two alleles

Here we consider the simple, special case of two alleles with possibly asymmetric mutation. Call the two alleles '1' and '2' and let θ_{12} and θ_{21} be the mutation rates from allele 1 to allele 2 and from allele 2 to allele 1, respectively, at each site. This general two-allele model can be converted into the PIM model with $K = 2$ by setting

$$\theta_L = \theta_{12}+\theta_{21}, \qquad \pi_1 = \theta_{21}/(\theta_{12}+\theta_{21}), \qquad \pi_2 = 1-\pi_1.$$

Then, we may apply the results of the previous subsection. Recalling that $\theta = L\theta_L$, then k^* is Poisson distributed in the limit with mean

$$\begin{aligned} \theta\sum_{j=1}^{[n/2]}c_j &= L(\theta_{12}+\theta_{21})\frac{2\theta_{12}\theta_{21}}{(\theta_{12}+\theta_{21})^2}\sum_{j=1}^{n-1}1/j \\ &= L\frac{2\theta_{12}\theta_{21}}{(\theta_{12}+\theta_{21})}\sum_{j=1}^{n-1}1/j, \end{aligned} \tag{17}$$

so that the overall mutation rate estimated using Watterson's estimator is equal to L times the harmonic mean of the two rates of mutation. If mutation is symmetric, and $\theta_{12} = \theta_{21} \equiv \theta^*$, then Watterson's estimator estimates $L\theta^*$.

2.4. Dependent sites

In this subsection we discuss the case of dependent sites, that is sites without free recombination between them. Note that, dependence (or lack thereof) between a set of random variables does not change the expected values. Therefore, any method-of-moments estimator based on the limiting distribution of independent sites remains valid for dependent sites, meaning that

its expected value is equal to the corresponding parameter. More rigorously, suppose that $Y_j^{(i)}$ is the indicator variable that site i has exactly two alleles with counts j and $n-j$. Then $k_j^* = \sum_{i=1}^{L} Y_j^{(i)}$, and

$$E(k_j^*) = E\left(\sum_{i=1}^{L} Y_j^{(i)}\right) = \sum_{i=1}^{L} E\left(Y_j^{(i)}\right),$$

which shows that the expected number of segregating sites depends only on the marginal expectation of $Y_j^{(i)}$. Also,

$$E(k^*) = \sum_{j=1}^{[n/2]} E(k_j^*) \to \theta \sum_{j=1}^{[n/2]} c_j.$$

All the estimators of θ described above are method-of-moments estimators, in addition to being MLEs of θ when sites are independent. Therefore, they will be valid as unbiased, method-of-moments estimators in the case of dependent sites.

3. Discussion

In this article we have proved that the number of sites segregating among a large number of independent sites is sufficient for estimating the mutation parameter θ. Our results show that the common interpretation of this parameter corresponds to particular highly symmetric models of mutation, when the mutation rate at each site is very small and there is a very large number of sites. Two examples are the K-allele model with symmetric mutation and the infinitely-many-alleles model under the condition in Eq. (16). In these cases, Watterson's estimator is the MLE of θ if sites are independent, and is an unbiased method-of-moments estimator of θ if sites are not independent.

Watterson's estimator is often applied to DNA sequence data. Although such data typically contain non-independent sites, some general conclusions of our work still apply. In particular, it is well known that mutation rates among the four nucleotides are not symmetric; e.g., see Chapter 13 of Felsenstein (2004). Although we were unable to obtain results for general mutation models with more than two alleles, the result for a general two-allele model shown in Eq. (17) and the result for a K-allele model with parent-independent in Eq. (15) demonstrate that Watterson's estimator estimates a net mutation rate which depends on the details of how mutation operates at individual sites. It is logical to re-define θ as this net mutation rate (θ_{net}). For example, $\theta_{\text{net}} = \theta(1 - \sum_{i=1}^{K} \pi_i^2)$ for the PIM model; for the symmetric mutation model $\theta_{\text{net}} = \theta(1 - 1/K)$; for the two-allele model

$$\theta_{\text{net}} = \frac{2\theta_{12}\theta_{21}}{(\theta_{12} + \theta_{21})},$$

the harmonic mean of the two mutation rates. It is, however, important to keep in mind that θ_{net} will then be a function of the mutation model.

Acknowledgments

We are grateful to Warren J. Ewens for inspiring this research and for many helpful comments. We also thank the reviewers for their constructive suggestions.

References

Beaumont, M.A., Zhang, W., Balding, D.J., 2002. Approximate bayesian computation in population genetics. Genetics 162, 2025–2035.
Bustamante, C.D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M.T., Glanowski, S., Hernandez, R., Civello, D., Tanenbaum, D.M., White, T.J., Sninsky, J.J., Adams, M.D., Cargill, M., Clark, A.G., 2005. The cost of inbreeding in Arabidopsis. Nature 437, 1153–1157.
Bustamante, C.D., Nielsen, R., Sawyer, S.A., Olsen, K.M., Purugganan, M.D., Hartl, D.L., 2002. The cost of inbreeding in Arabidopsis. Nature 416, 531–534.
Desai, M., Plotkin, J.B., 2008. The polymorphism frequency spectrum of finitely many sites under selection. Genetics 180, 2175–2191.
Ewens, W.J., 1982. On the concept of effective size. Theor. Pop. Biol. 21, 373–378.
Ewens, W.J., 1974. A note on the sampling theory for infinite alleles and infinite sites models. Theor. Pop. Biol. 6, 143–148.
Ewens, W.J., 1972. The sampling theory of selectively neutral alleles. Theor. Pop. Biol. 3, 87–112.
Feller, W., 1970. An Introduction to Probability Theory and Its Applications, third ed.. vol. 1. Wiley, New York.
Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associates. Sunderland, Massachusetts.
Fisher, R.A., 1930. The Genetical Theory of Natural Selection. Clarendon Press, Oxford.
Fu, X., 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147, 915–925.
Hudson, R.R., 1983. Properties of a neutral allele model with intragenic recombination. Theor. Pop. Biol. 23, 183–201.
Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. Mamm. Protein Metabol. 21–32.
Kaplan, N.L., Hudson, R.R., 1985. The use of sample genealogies for studying a selectively neutral m-loci model with recombination. Theor. Pop. Biol. 28, 382–396.
Kimura, M., 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61, 893–903.
Kimura, M., Crow, J.F., 1964. The number of alleles that can be maintained in a finite population. Genetics 49, 725–738.
Malécot, G., 1946. La consaguinité dans une population limitée. C. R. Acad. Sci. Paris 222, 841–843.
Sawyer, S.A., Hartl, D.L., 1992. Population genetics of polymorphism and divergence. Genetics 132, 1161–1176.
Sawyer, S.A., Kulathinal, R.J., Bustamante, C.D., Hartl, D.L., 2003. Bayesian analysis suggests that most amino acid replacements in drosophila are driven by positive selection. J. Mol. Evol. 57, S154–S164.
Sjödin, P., Kaj, I., Krone, S., Lascoux, M., Nordborg, M., 2005. On the meaning and existence of an effective population size. Genetics 169, 1061–1070.
Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585–595.
Wakeley, J., 2003. Polymorphism and divergence for island-model species. Genetics 163, 411–420.
Watterson, G.A., 1975. On the number of segregating sites in genetic models without recombination. Theor. Pop. Biol. 7, 256–276.
Williamson, S., Fledel-Alon, A., Bustamante, C.D., 2004. Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. Genetics 168, 463–475.
Wright, S., 1931. Evolution in Mendelian populations. Genetics 16, 97–159.
Wright, S., 1949. Adaptation and selection. In: Jepson, G.L., Simpson, G.G., Mayr, E. (Eds.), Genetics, Paleontology and Evolution. Princeton Univ. Press, Princeton.
Zhu, L., Bustamante, C.D., 2005. A composite-likelihood approach for detecting directional selection from DNA sequence data. Genetics 170, 1411–1421.