Theoretical Population Biology 102 (2015) 94–101

Contents lists available at ScienceDirect

Theoretical Population Biology

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

Enhancing the mathematical properties of new haplotype homozygosity statistics for the detection of selective sweeps

Nandita R. Garud [a,*], Noah A. Rosenberg [b,*]

[a] Department of Genetics, Stanford University, Stanford, CA 94305, USA
[b] Department of Biology, Stanford University, Stanford, CA 94305, USA

ARTICLE INFO

Article history:
Received 27 March 2015
Available online 16 April 2015

Keywords:
Haplotype statistics
Selective sweeps
Drosophila melanogaster

ABSTRACT

Soft selective sweeps represent an important form of adaptation in which multiple haplotypes bearing adaptive alleles rise to high frequency. Most statistical methods for detecting selective sweeps from genetic polymorphism data, however, have focused on identifying hard selective sweeps in which a favored allele appears on a single haplotypic background; these methods might be underpowered to detect soft sweeps. Among exceptions is the set of haplotype homozygosity statistics introduced for the detection of soft sweeps by Garud et al. (2015). These statistics, examining frequencies of multiple haplotypes in relation to each other, include H_{12}, a statistic designed to identify both hard and soft selective sweeps, and H_2/H_1, a statistic that conditional on high H_{12} values seeks to distinguish between hard and soft sweeps. A challenge in the use of H_2/H_1 is that its range depends on the associated value of H_{12}, so that equal H_2/H_1 values might provide different levels of support for a soft sweep model at different values of H_{12}. Here, we enhance the H_{12} and H_2/H_1 haplotype homozygosity statistics for selective sweep detection by deriving the upper bound on H_2/H_1 as a function of H_{12}, thereby generating a statistic that normalizes H_2/H_1 to lie between 0 and 1. Through a reanalysis of resequencing data from inbred lines of *Drosophila*, we show that the enhanced statistic both strengthens interpretations obtained with the unnormalized statistic and leads to empirical insights that are less readily apparent without the normalization.

1. Introduction

A selective sweep, the process whereby beneficial mutations at a locus that contribute to the fitness of an organism rise in frequency to become prevalent in a population, can occur through two main mechanisms that leave distinct genomic signatures (Pritchard et al., 2010; Cutter and Payseur, 2013; Messer and Petrov, 2013). A relatively new adaptive allele can proliferate so that the single haplotype on which it has occurred reaches a high frequency, resulting in a signature of a "hard" selective sweep (Maynard Smith and Haigh, 1974; Kaplan et al., 1989; Kim and Stephan, 2002). Alternatively, a mutation that arises *de novo* multiple times or exists as standing genetic variation on several haplotype backgrounds before the onset of positive selection can increase in frequency; in these cases, multiple favored haplotypes have relatively high frequencies, generating a signature of a "soft" selective sweep (Hermisson and Pennings, 2005; Przeworski et al., 2005; Pennings and Hermisson, 2006a). Soft sweeps can provide an effective mechanism for natural selection and might explain a sizeable fraction of selective events in many systems (Orr and Betancourt, 2001; Innan and Kim, 2004; Pritchard et al., 2010; Messer and Petrov, 2013).

Most statistical methods that have been designed to detect selective sweeps from patterns of genetic polymorphism search for patterns expected under a hard-sweep model, such as the presence of a single common haplotype (Hudson et al., 1994), high haplotype homozygosity (Depaulis and Veuille, 1998; Sabeti et al., 2002; Voight et al., 2006), high-frequency derived variants and related features of site-frequency spectra (Tajima, 1989; Braverman et al., 1995; Fay and Wu, 2000; Nielsen et al., 2005), or local loss of variation near a putative selected site (Maynard Smith and Haigh, 1974; Begun and Aquadro, 1992; Kim and Stephan, 2002). Many methods that search for patterns expected with hard sweeps, however, can be less well suited to the problem of identifying soft sweeps (Pennings and Hermisson, 2006b; Teshima et al., 2006; Cutter and Payseur, 2013). Therefore, current genomic scans for selective sweeps might be limited in their ability to uncover an important class of adaptive events.

* Corresponding authors.
E-mail addresses: ngarud@stanford.edu (N.R. Garud), noahr@stanford.edu (N.A. Rosenberg).

http://dx.doi.org/10.1016/j.tpb.2015.04.001
0040-5809/

Recently, it has been shown that statistics based on haplotype homozygosity can identify both hard and soft sweeps from population-genomic data (Ferrer-Admetlla et al., 2014; Garud et al., 2015). Garud et al. (2015) developed a haplotype homozygosity statistic, H_{12}, relying on the principle that in a soft sweep, the most frequent haplotype might not predominate in frequency, and instead, multiple frequent haplotypes might be present. In terms of frequencies $p_i \geq 0$ for $i = 1, 2, 3, \ldots$ with $\sum_{i=1}^{\infty} p_i = 1$ and $p_1 \geq p_2 \geq p_3 \geq \cdots$, Garud et al. (2015) defined H_{12} as

$$H_{12} = (p_1 + p_2)^2 + \sum_{i=3}^{\infty} p_i^2. \tag{1}$$

This statistic calculates homozygosity by combining the two largest haplotype frequencies into a single value and then computing a haplotype homozygosity. Garud et al. (2015) determined that H_{12} has reasonable power to detect both hard and soft sweeps, applying the statistic to *Drosophila* population-genomic data and identifying abundant signatures of natural selection.

To determine whether the genomic regions with the highest values of H_{12} were compatible with either a hard-sweep or soft-sweep pattern, Garud et al. (2015) examined a second statistic, H_2/H_1, a ratio of a haplotype homozygosity H_2 that excludes the most frequent haplotype and a haplotype homozygosity H_1 that includes this haplotype:

$$H_1 = p_1^2 + p_2^2 + \sum_{i=3}^{\infty} p_i^2 \tag{2}$$

$$H_2 = p_2^2 + \sum_{i=3}^{\infty} p_i^2. \tag{3}$$

For high values of H_{12}, hard sweeps are expected to produce relatively low values of H_2/H_1 because they produce a single high-frequency haplotype (very high p_1, low p_2). Soft sweeps, on the other hand, produce multiple high-frequency haplotypes (high p_1, p_2, and perhaps others), and are expected to produce higher values of H_2/H_1.

Garud et al. (2015) found that this two-step process – identification of regions with high H_{12} followed by examination of H_2/H_1 – could both detect selective sweeps in general and distinguish hard and soft sweeps. As we will show, however, a complication in the approach is that the permissible range of H_2/H_1 varies with the value of H_{12}. Thus, the magnitude of H_2/H_1 that might be regarded as indicative of a soft or hard sweep can depend on the associated values of H_{12}. This potential difference in interpretations for values of H_2/H_1 as a function of H_{12} can present a particular challenge when comparing H_2/H_1 at multiple loci with a wide range of H_{12} values.

In a line of work separate from the use by Garud et al. (2015) of homozygosity-based soft sweep statistics, Rosenberg and Jakobsson (2008) and Reddy and Rosenberg (2012) analyzed the properties of homozygosity statistics in relation to the frequency of the most frequent allele, identifying upper and lower bounds on homozygosity given the frequency of the most frequent allele. This work, along with related work on other statistics (Long and Kittles, 2003; Hedrick, 2005; Jost, 2008; VanLiere and Rosenberg, 2008; Maruki et al., 2012; Jakobsson et al., 2013), seeks to understand mathematical bounds on population-genetic statistics, so that their application and interpretation can be suitably informed by the mathematical constraints on their numerical values.

Here, to facilitate the interpretation of the statistics of Garud et al. (2015) and to enhance comparisons among values of these statistics at loci with different haplotype homozygosities, we use a result from Rosenberg and Jakobsson (2008) to determine the upper and lower bounds on H_2/H_1 as a function of H_{12}. The upper bound provides a basis for normalization of H_2/H_1 to produce a statistic with the same range, from 0 to 1, irrespective of the value of H_{12}. Using the upper bound and the new normalized statistic, we reexamine *Drosophila* data analyzed by Garud et al. (2015), demonstrating that the upper bound, $(H_2/H_1)_{\max}$, and the normalized statistic, $(H_2/H_1)'$, enable improved insights regarding soft selective sweeps on the basis of genetic polymorphism data.

2. Theory

Our goal is to determine the maximum of H_2/H_1 given the value of H_{12}, for $0 < H_{12} \leq 1$. For convenience, we denote $Z = H_2/H_1$. We denote the desired upper bound by $Z_{\max}$.

For generality in our description, we consider "alleles" at a locus. These distinct "alleles" can be viewed as representing distinct haplotypes at a specific location in the genome; the assumption is that a set of distinct genetic types is considered, representing perhaps distinct haplotypes or distinct alleles in the traditional sense, and the sum of the frequencies of the types is 1.

We sort alleles in descending order of frequency, so that $p_1 > 0$ and $p_1 \geq p_2 \geq p_3 \geq \cdots \geq 0$. The number of alleles is left unspecified, and it can be arbitrarily large; thus, $\sum_{i=1}^{\infty} p_i = 1$. For our mathematical analysis, we consider parametric allele frequencies; that is, the p_i are treated as known frequencies in a population rather than values estimated from samples. The mathematical setting follows Rosenberg and Jakobsson (2008).

We let $M = p_1 + p_2$. Because $p_1 > 0$, M, H_{12}, and H_1 are all strictly positive. By analogy with H_1 and H_2, denote $H_3 = \sum_{i=3}^{\infty} p_i^2$. Thus, by Eq. (1),

$$H_{12} = M^2 + H_3. \tag{4}$$

2.1. The upper bound on H_2/H_1 given H_{12}

We proceed in two main steps. First, for fixed H_{12} and fixed M, we determine the maximum of Z as a function of p_1. Next, we identify the value of M that maximizes Z. This pair of steps constructs the set of allele frequencies $\{p_i\}_{i=1}^{\infty}$ that generates the maximal Z at fixed H_{12}. A graphical overview of the argument appears in Fig. 1.

Maximizing Z for fixed H_{12} and fixed M. Because $H_2 = p_2^2 + H_3$ and $p_2 = M - p_1$, H_2 can be rewritten

$$H_2 = (M - p_1)^2 + H_3. \tag{5}$$

Note that by Eq. (4), for fixed H_{12} and fixed M, H_3 is constant. Because $M = p_1 + p_2$, $p_1 \geq p_2$, and $p_1 > 0$, it follows that $M/2 \leq p_1 \leq M$. Treated as a function of p_1, on the interval $[M/2, M]$, $(M - p_1)^2 + H_3$ is decreasing.

Using Eq. (5), $Z = H_2/H_1$ can be written as

$$Z = \frac{(M - p_1)^2 + H_3}{p_1^2 + (M - p_1)^2 + H_3} = \frac{1}{p_1^2/[(M - p_1)^2 + H_3] + 1}. \tag{6}$$

In Eq. (6), for fixed H_{12} and fixed M, p_1^2 is increasing in p_1 and $(M - p_1)^2 + H_3$ is decreasing. The ratio $p_1^2/[(M - p_1)^2 + H_3]$ is therefore increasing in p_1, so that the entire expression for Z decreases with p_1. It is therefore maximized when p_1 is minimized—in other words, when $p_1 = p_2 = M/2$. The maximal Z for fixed H_{12} and fixed M is

$$Z = \frac{4H_{12} - 3M^2}{4H_{12} - 2M^2}. \tag{7}$$

It remains to maximize Z by finding the value of M that maximizes Eq. (7) for fixed H_{12}. By rewriting Eq. (7) as $Z = 1 - M^2/(4H_{12} - 2M^2)$, it can be seen that for fixed H_{12}, as M increases, M^2