Published in final edited form as:

Theor Popul Biol. 2015 June ; 102: 94–101. doi:10.1016/j.tpb.2015.04.001.

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

Nandita R. Garud[*] and **Noah A. Rosenberg**[†]

[*] Department of Genetics, Stanford University, Stanford, CA 94305 USA

[†] Department of Biology, Stanford University, Stanford, CA 94305 USA

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.

Keywords

Haplotype statistics; selective sweeps; *Drosophila melanogaster*

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

ngarud@stanford and noahr@stanford.edu

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

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.

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 $p_1 \geq p_2 \geq p_3 \geq \ldots$, GARUD *et al.* (2015) defined H_{12} as

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

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:

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

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

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.

Theory

Our goal is to determine the maximum of H_2/H_1 given the value of H_{12}, for $0 < H_{12} \le 1$. 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 $p_1 \geq p_2 \geq p_3 \geq \ldots \geq 0$. 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,

$$H_{12} = M^2 + H_3. \quad (4)$$

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 Figure 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

$$H_2 = (M - p_1)^2 + H_3. \quad (5)$$

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

$$\begin{aligned} Z &= \frac{(M-p_1)^2 + H_3}{p_1^2 + (M-p_1)^2 + H_3} \\ &= \frac{1}{p_1^2 / \left[(M-p_1)^2 + H_3\right] + 1}. \end{aligned} \quad (6)$$

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 $p_1^2 / \left[(M - p_1)^2 + H_3\right]$ 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

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

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 increases, $4H_{12} - 2M^2$ decreases, and Z decreases. Thus, for fixed H_{12}, the maximal Z, treated as a function of M, occurs when M isas small as possible.

The minimal value of *M* given *H_{12}*—We have shown that maximizing Z for fixed H_{12} and M requires $p_1 = p_2 = M/2$, and hence, using the descending order of the allele frequencies, $p_3 \le M/2$. We have also shown that maximizing Z for fixed H_{12} over all possible M requires us to find the minimal M permissible for fixed H_{12}. This problem can be solved with a known result. We first ignore the trivial case of $H_{12} = 1$, for which the maximal Z has $M = 1$, $p_1 = p_2 = 1/2$, $H_1 = 1/2$, $H_2 = 1/4$, and $Z_{\max} = 1/2$.

By eq. 4, minimizing M for fixed H_{12} amounts to maximizing H_3. Lemma 3 of ROSENBERG and JAKOBSSON (2008) obtains the maximal sum of squares for a set of nonnegative numbers in a non-increasing sequence, each of which lies below the same specified constant, and whose sum is specified. In our case, the sequence is $\{p_i\}_{i=3}^{\infty}$, the entries are bounded above by $M/2$, and their sum is $1 - p_1 - p_2 = 1 - M$.

Applying the lemma, we obtain

$$H_3 \le K(K-1)\left(\frac{M}{2}\right)^2 - 2(1-M)(K-1)\frac{M}{2} + (1-M)^2, \quad (8)$$

where $K = \lceil (1 - M)/(M/2) \rceil = \lceil 2/M \rceil - 2$ and $\lceil x \rceil$ denotes the smallest integer greater than or equal to x; in the application of the lemma, K gives the number of nonzero numbers in the sequence $\{p_i\}_{i=3}^{\infty}$ that achieves the maximum. Equality is achieved if and only if $\lceil 2/M \rceil - 3$ alleles (in addition to alleles 1 and 2) have frequency $M/2$, and one allele has frequency $(1 - M) - (\lceil 2/M \rceil - 3)(M/2) = 1 - (\lceil 2/M \rceil - 1)(M/2)$.

The minimal M is obtained by substituting the upper bound from eq. 8 for H_3 in eq. 4 and solving for M. The equation that must be solved is

$$H_{12} = \frac{K^2 + 3K + 4}{4} M^2 - (K+1)M + 1. \quad (9)$$

Note that K is currently considered a function of M, equaling $\lceil 2/M \rceil - 2$. However, we can instead determine the value of K as a function of H_{12}, so that eq. 9 becomes a simple quadratic equation in M. To solve eq. 9 for M at a given H_{12}, we must find the value of K—the number of alleles of nonzero frequency (not including alleles 1 and 2)—that applies for the given value of H_{12}.

We break the unit interval (0, 1) into disjoint intervals $[2/I, 2/(I-1))$ for integers $I \geq 3$. On the interval $[2/I, 2/(I-1))$ for M, $K = I - 2$. Inserting $K = I - 2$ into eq. 9, for M in this interval, the minimal M in terms of H_{12} is obtained by solving

$$H_{12}=\frac{I^2-I+2}{4}M^2-(I-1)M+1 \quad (10)$$

for M. Thus, identifying the value of K in terms of H_{12} for use in eq. 9 amounts to finding the value of I in terms of H_{12} for use in eq. 10.

The right-hand side of eq. 10 is monotonically increasing on the interval $[2/I, 2/(I-1))$, as it is a concave-up parabola in M with minimum at $M = 2(I-1)/(I^2 - I + 2) = 2/[I + 2/(I-1)] < 2/I$. The vertex of the parabola lies to the left of the left endpoint of the interval, $M = 2/I$, so that on $[2/I, 2/(I-1))$, the parabola is increasing.

At the left endpoint $M = 2/I$, $H_{12} = (I + 2)/I^2$, and at the right endpoint $M = 2/(I-1)$, $H_{12} = (I + 1)/(I-1)^2$. Consequently, because H_{12} increases as a function of M on the interval $[2/I, 2/(I-1))$, for this interval, H_{12} lies in $[(I + 2)/I^2, (I + 1)/(I-1)^2)$.

As a strictly monotonic continuous function from $[2/I, 2/(I-1))$ to $[(I + 2)/I^2, (I + 1)/(I-1)^2)$, H_{12} is invertible as a function of M. Treated as a function of M in (0, 1), I satisfies $2/I \leq M < 2/(I-1)$; similarly, as a function of H_{12} in (0, 1), I satisfies $(I + 2)/I^2 \leq H_{12} < (I + 1)/(I-1)^2$. In other words, given H_{12}, I must be equal to the smallest integer for which $(I + 2)/I^2 \leq H_{12}$.

Solving this inequality, either $I \geq \left(1+\sqrt{8H_{12}+1}\right)/(2H_{12})$ or $I \leq \left(1-\sqrt{8H_{12}+1}\right)/(2H_{12})$. The latter root is negative and can be discarded as $I \geq 3$. The smallest integer that satisfies the former inequality is

$$I=\lceil\frac{1+\sqrt{8H_{12}+1}}{2H_{12}}\rceil. \quad (11)$$

We can now complete the solution for the minimal M as a function of H_{12}: this minimum is a solution to eq. 10 when eq. 11 is used for I. The equation has two roots; the smaller root is smaller than $2/I$, and therefore lies outside the interval $[2/I, 2/(I-1))$ in which M must fall when H_{12} satisfies eq. 11. The minimal M therefore equals the larger root.

The formula for Z_{max}—Compiling the steps we have completed, we have that as a function of H_{12},

$$Z_{max}(H_{12})=\frac{4H_{12}-3M^2}{4H_{12}-2M^2}, \quad (12)$$

where M is the larger root of eq. 10,

$$M = \frac{2\left(I-1\right)+2\sqrt{\left(I^2-I+2\right)H_{12}-\left(I+1\right)}}{I^2-I+2}, \quad (13)$$

and I satisfies eq. 11. The formula for $Z_{\max}$ holds for all H_{12} in (0, 1]; in the $H_{12} = 1$ case that we initially discarded, eq. 12 gives the correct value $Z_{\max} = 1/2$. $Z_{\max}$ is reached when $I - 1$ alleles each have frequency $M/2$ and one allele has frequency $1 - (\lceil 2/M \rceil - 1)(M/2)$.

Figure 2 plots eq. 12 as a function of H_{12} over the unit interval. A piecewise structure of the upper bound $Z_{\max}$ is visible, reflecting the fact that at points $H_{12} = (I + 2)/I^2$ for integers $I \geq 3$, the value of I as a function of H_{12} changes, and $Z_{\max}$ is not differentiable. $Z_{\max}$ approaches a limiting value of 1 as H_{12} approaches 0, and it declines monotonically to a value of 1/2 at $H_{12} = 1$.

An approximation to $Z_{\max}$—It is convenient to consider a simple approximation to $Z_{\max}$ by examining the points $H_{12} = (I + 2)/I^2$ for integers $I \geq 3$. At these points, applying eqs. 11-13,

$$I = \frac{1+\sqrt{8H_{12}+1}}{2H_{12}}, \quad (14)$$

$M = 2/I$, and $Z_{\max} = (I - 1)/I$. Eqs. 11-13 simplify because $Z_{\max}$ is achieved when I alleles each have frequency $1/I$, unlike for other $H_{12} < 1$, where one nonzero frequency differs from the others.

We can approximate $Z_{\max}$ by finding a function $Y_{\max}$ that satisfies

$$Y_{max}\left(\frac{I+2}{I^2}\right) = \frac{I-1}{I}$$

at the points specified by integers $I \geq 3$ and using this function to interpolate across all values of H_{12}. When $H_{12} = (I + 2)/I^2$ for integers $I \geq 3$, I satisfies eq. 14, and

$$\frac{I-1}{I} = \frac{1+\sqrt{8H_{12}+1}-2H_{12}}{1+\sqrt{8H_{12}+1}}.$$

Multiplying the numerator and denominator of this equation by $1-\sqrt{8H_{12}+1}$, we have

$$Y_{max}\left(H_{12}\right) = \frac{5-\sqrt{8H_{12}+1}}{4}. \quad (15)$$

This approximate bound agrees with the strict bound $Z_{\max}$ at points $H_{12} = (I + 2)/I^2$ for integers $I \geq 3$, and it matches $Z_{\max}$ at the endpoints of the unit interval. In Figure 2, it can be seen that $Y_{\max}$ provides a reasonable approximation to $Z_{\max}$ over the entire interval.

Not only is $Y_{\max}$ an approximation to the strict upper bound $Z_{\max}$, $Y_{\max} \geq Z_{\max}$ for H_{12} in (0, 1], so that $Y_{\max}$ is itself an upper bound. To prove this result, using eqs. 15 and 12, we have

$$Y_{max}\left(H_{12}\right)-Z_{max}\left(H_{12}\right)=\frac{\left(2H_{12}+M^2\right)-\left(2H_{12}-M^2\right)\sqrt{8H_{12}+1}}{4\left(2H_{12}-M^2\right)}.$$

The denominator is positive, as $2H_{12} - M^2 = M^2 + 2H_3 > 0$. It remains to show that

$$2H_{12}+M^2\geq\left(2H_{12}-M^2\right)\sqrt{8H_{12}+1}.$$

Squaring both sides, $Y_{\max}(H_{12}) - Z_{\max}(H_{12}) \geq 0$ if

$$-8H_{12}\left[4H_{12}^2-4M^2H_{12}+\left(M^4-M^2\right)\right]\geq 0.$$

As H_{12} is positive, $Y_{\max}(H_{12}) - Z_{\max}(H_{12}) \geq 0$ if H_{12} lies in the closed interval bounded by the roots of the quadratic term in brackets, or $(M^2 - M)/2$ and $(M^2 + M)/2$. Because $0 < M \leq 1$, the smaller root is at most 0, and $H_{12} \geq (M^2 - M)/2$ always holds. It thus suffices to prove $H_{12} \leq (M^2 + M)/2$.

Recalling eq. 4, we must show that $H_3 \leq (M - M^2)/2$. Eq. 8 provides a maximum on H_3 in terms of M; substituting this maximum for H_3, we have $H_3 \leq (M - M^2)/2$ if

$$\frac{1}{4}\left(KM+M-2\right)\left(KM+2M-2\right)\leq 0.$$

This last inequality is true by definition of $K = \lceil 2/M \rceil - 2$, as $2/M - 2 \leq K < 2/M - 1$ implies $KM + M - 2 < 0$ and $KM + 2M - 2 \geq 0$. We can therefore conclude that $H_3 \leq (M - M^2)/2$, and hence $H_{12} \leq (M^2 + M)/2$, and $Y_{\max}(H_{12}) - Z_{\max}(H_{12}) \geq 0$ for H_{12} in (0, 1].

The lower bound on H_2/H_1 given H_{12}

It is straightforward to show that for any H_{12} in (0, 1], H_2/H_1 can get arbitrarily close to 0. For $H_{12} = 1$, we set $p_1 = 1 - \varepsilon$ and $p_2 = \varepsilon$ for a small $\varepsilon > 0$. Then $H_2/H_1 = \varepsilon^2/[(1 - \varepsilon)^2 + \varepsilon^2]$, which approaches 0 as $\varepsilon \to 0$. Otherwise, we construct a scenario with one frequent allele and K rare alleles, and demonstrate that $H_2/H_1 \to 0$ as $K \to \infty$.

Suppose $p_1=\sqrt{KH_{12}-1}/\sqrt{K-1}$ and $p_2=p_3=\ldots=p_{K+1}=\left(1-\sqrt{KH_{12}-1}/\sqrt{K-1}\right)/K$ for large K. Frequency p_1 is large and the remaining frequencies are small. In this case,

$$\begin{aligned}\frac{H_2}{H_1}= & \frac{Kp_2^2}{p_1^2+Kp_2^2}\\ = & \frac{\left(1-\frac{\sqrt{KH_{12}-1}}{\sqrt{K-1}}\right)^2/K}{\frac{KH_{12}-1}{K-1}+\left(1-\frac{\sqrt{KH_{12}-1}}{\sqrt{K-1}}\right)^2/K}\\ = & \frac{\left(\sqrt{K-1}-\sqrt{KH_{12}-1}\right)^2}{K(KH_{12}-1)+\left(\sqrt{K-1}-\sqrt{KH_{12}-1}\right)^2}.\end{aligned}$$

The denominator has higher degree in K than the numerator, so that $\lim_{K \to \infty}(H_2/H_1) = 0$.

The mean range of H_2/H_1 given H_{12}

Determining the mean of the range of Z, treated as a function of H_{12} over the unit interval, can provide a sense of the magnitude of the constraint placed by H_{12} on Z. For a statistic with a larger mean range, a greater proportion of the unit interval can be achieved, and the statistic is less constrained than is one with a smaller mean range.

Because $Y_{\max}(H_{12}) \geq Z_{\max}(H_{12})$, the simpler $Y_{\max}$ can assist in evaluating the mean size of the range of Z. As the minimum Z approaches 0 for all H_{12} in (0, 1], the size of the range for Z is simply $Z_{\max}$. On the unit interval, $Y_{\max}$ has mean

$$\int_0^1 Y_{max}\left(H_{12}\right) dH_{12} = \frac{17}{24} \approx 0.708, \quad (16)$$

and therefore, the mean $Z_{\max}$ is smaller than 17/24. This mean exceeds 1/2, as the minimal $Z_{\max}$ for H_{12} in (0, 1], at $H_{12} = 1$, is 1/2. Numerical integration of eq. 12 to obtain the mean $Z_{\max}$ gives

$$\sum_{I=3}^{\infty} \int_{(I+2)/I^2}^{(I+1)/(I-1)^2} Z_{max}\left(H_{12}\right) dH_{12} \approx 0.684. \quad (17)$$

This result illustrates that the mean across the unit interval for H_{12} of the error in the approximation of $Z_{\max}$ by $Y_{\max}$ is small, approximately 0.708 – 0.684 = 0.024. Further, although the range of Z is constrained, the mean range over all values of H_{12} in (0, 1] is larger than corresponding mean constraints in other contexts involving homozygosity, F_{st}, the r^2 statistic for linkage disequilibrium, and the frequency of the most frequent allele (ROSENBERG *et al.*, 2003; ROSENBERG and JAKOBSSON, 2008; VANLIERE and ROSENBERG, 2008; REDDY and ROSENBERG, 2012; JAKOBSSON *et al.*, 2013; EDGE and ROSENBERG, 2014).

Normalized statistics

Because H_2/H_1 can approach 0 for any H_{12}, a normalization of H_2/H_1 to lie in [0, 1] need only be concerned with the upper bound on H_2/H_1. We can therefore define exact and approximate normalizations of Z at given values of H_{12} as follows:

$$Z' = \frac{Z}{Z_{max}\left(H_{12}\right)} \quad (18)$$

$$Z'' = \frac{Z}{Y_{max}\left(H_{12}\right)}, \quad (19)$$

The denominators of these equations are computed using eqs. 12 and 15, respectively.

Application to data

We illustrate the bounds on H_2/H_1 as functions of H_{12} by reexamining two *Drosophila melanogaster* data sets studied by GARUD *et al.* (2015), each containing fully sequenced genomes of inbred lines generated from samples taken in North Carolina. First, we consider the *Drosophila* Genetic Reference Panel (DGRP) data set consisting of sequences of 145 inbred lines (MACKAY *et al.*, 2012). Next, we examine the *Drosophila* Population Genomic Panel (DPGP) consisting of 40 strains. We consider these two data sets generated with different samples both to show an example use of the upper bounds and to demonstrate how inferences from samples with differing numerical patterns in H_{12} and H_2/H_1 can be viewed as comparable.

DGRP data

We first consider the DGRP data set studied by GARUD *et al.* (2015). As a consequence of inbreeding, the DGRP genomes are largely homozygous. On each of the four autosomal arms, GARUD *et al.* (2015) examined haplotypes within analysis windows of 400 single-nucleotide polymorphisms (SNPs, ~10kb). Because low recombination rates can result in high haplotype homozygosities, GARUD *et al.* (2015) excluded analysis windows overlapping 100 kb tracts measured by COMERON *et al.* (2012) to have recombination rates lower than 5 × 10^{-7} centimorgans per base pair (cM/bp). To classify haplotypes within windows, GARUD *et al.* (2015) assigned the 400-SNP haplotypes into groups according to exact sequence identity. If a haplotype with missing data matched multiple haplotypes at all genotyped sites in the analysis window, then the haplotype was randomly assigned to one of these classes. In the DGRP data set, all heterozygous sites in a strain were treated as missing data. Examining all 4,013,703 segregating sites across the 145 strains, 0.7% heterozygous sites per base pair per strain and 4.2% missing data per base pair per strain were observed. If a haplotype could not be conclusively assigned based on the information at non-missing data sites, then the haplotype was randomly assigned to a haplotype class that matched at all other sites; across all analysis windows and strains, 18% of assignments to haplotype classes used this method of random assignment. Windows were incremented by 50 SNPs, so that consecutive windows overlapped by 350 SNPs.

Each window has a haplotype frequency distribution across the 145 lines, enabling computations of H_{12}, H_1, and H_2. To avoid inflating the number of selective events inferred in a genomic region, GARUD *et al.* (2015) grouped together consecutive windows as belonging to the same "peak" if the H_{12} values in all of the grouped windows were above a critical H_{12} value calculated under a neutral demographic model. They assigned H_{12} and H_2/H_1 values to individual peaks by using the values calculated in the analysis window with the largest H_{12} within a peak. GARUD *et al.* (2015) focused on the 50 peaks with the largest H_{12} values, none of which possessed two or more windows sharing the same highest H_{12} value. The top three peaks coincided with the loci *Ace*, *Cyp6g1*, and *CHKov1*, prominent cases of adaptation previously discovered by detailed focused analyses (DABORN *et al.*, 2001; CATANIA *et al.*, 2004; MENOZZI *et al.*, 2004; AMINETZACH *et al.*, 2005; KARASOV *et al.*, 2010; SCHMIDT *et al.*, 2010; MAGWIRE *et al.*, 2011).

Effect of normalization in the DGRP data

We assessed the effect of the application of Z' to H_{12} and H_2/H_1 values calculated for the top 50 peaks in the DGRP data set. To do so across the full range of possible values for (H_{12}, H_2/H_1), we first calculated the change $\delta = Z' - Z$ in H_2/H_1 produced by the normalization.

For any value of H_{12}, as H_2/H_1 increases, δ also increases, reflecting the monotonicity of the upper bound on H_2/H_1 with increasing H_{12} (Figure 3A). The maximal δ of 1/2 is achieved when H_{12} = 1 and H_2/H_1 = 1/2.

Overlaid in Figure 3A are the H_{12} and H_2/H_1 values from the 50 top peaks in the DGRP data set. The values of H_{12} generally lie below 0.25, with most values occurring near 0.1. The values of H_2/H_1 span a wide range, with most (H_{12}, H_2/H_1) combinations lying in a region of the space where δ is between 0.025 and 0.05.

DPGP data

Our second example considers the DPGP data set that was also studied by GARUD *et al.* (2015). The DPGP data set (MACKAY *et al.*, 2012) consists of 40 of the original 145 inbred lines in the DGRP data set, sequenced and assembled separately from the DGRP data (www.dpgp.org).

In the DPGP data set, considering all 2,337,358 segregating sites across the 40 lines, there were 1.2% heterozygous sites per base pair per strain, and the missing data rate was 7.5%. With this reduced sample size compared to the DGRP data—and hence, with both shorter distances over which haplotypes become unique and faster computation times—GARUD *et al.* (2015) measured H_{12} values in shorter overlapping analysis windows of 100 SNPs incremented by 1 SNP. The treatment of haplotypes and missing data proceeded in the same manner as in the DGRP analysis. In this scan, averaging across lines, haplotypes with missing data were clustered with other haplotypes matching at all other positions at a lower rate of 2.7%.

As in the DGRP analysis, GARUD *et al.* (2015) identified the 50 peaks with the highest H_{12}. This analysis produced a distinct but overlapping set of high-H_{12} windows as the DGRP top 50 peaks, again recovering known cases of adaptation at *Ace*, *Cyp6g1*, and *CHKov1*.

Effect of normalization in the DPGP data

As in our analysis of the DPGP data, we assessed the effect of the application of Z' to high-H_{12} peaks in the DPGP data set. Figure 3B plots the (H_2/H_1, H_{12}) values for the top 50 peaks in the DPGP data. In comparison to those seen in the DGRP data set, the H_{12} values in the DPGP data are generally greater, and the H_2/H_1 values lower. As a consequence, the points in the DPGP data lie in a region of the space in which normalization has a greater effect, often with $\delta > 0.05$.

Comparison of DGRP and DPGP

GARUD *et al.* (2015) compared the positions of the top 50 peaks in the DPGP data set according to H_{12} with the positions of the top 50 peaks in the DGRP data set to determine if the same selective events were identified in the two data sets. To do so, GARUD *et al.* (2015) overlapped the edge coordinates of the peaks in the two data sets, where the edge coordinates of each peak correspond to the positions of the first SNP of the first analysis window and the last SNP of the last analysis window within a peak. An overlap was defined as a non-empty intersection of the two genomic regions defining the boundaries of the two peaks, one from one data set and one from the other. GARUD *et al.* (2015) found that 16 DPGP peaks overlapped

13 DGRP peaks, 10 of which were among the top 15 peaks in the DGRP scan. In three cases, two DPGP peaks overlapped one DGRP peak because multiple non-overlapping peaks in the DPGP data were in the same region as a DGRP peak. These multiple proximate peaks in the DPGP data set might have been part of the same selective events.

Jointly considering the DGRP and DPGP data sets, different sample depths and analysis window sizes can result in different distributions of H_{12} and H_2/H_1 values, and thus, in different inferences about selection. As a consequence, although several H_{12} peaks overlap in the DGRP and DPGP scans, the H_{12} and H_2/H_1 values for the top peaks differ between the two data sets. This result complicates the comparison of the selection signals obtained between the two data sets. Application of our normalization, however, can facilitate a meaningful comparison of the H_{12} and H_2/H_1 values measured in different data sets that potentially uncover the same selective events.

We applied the Z' and Z'' normalizations to overlapping peaks in the two data sets. Figure 4A shows that prior to normalization, the H_2/H_1 values for DGRP exceed those of DPGP, as was seen previously in the plots of all 50 windows in Figure 3. However, after normalization, the distributions of H_2/H_1 values for the two scans are comparable despite the differences in H_{12}. We quantified this change with a paired two-tailed Wilcoxon signed-rank test, testing the null hypothesis that the distributions of H_2/H_1 values in the DGRP and DPGP data are the same before and after application of Z' and Z''. Because 16 peaks in the DPGP data set overlap 13 peaks in the DGRP set, where three pairs of DPGP peaks each overlap unique peaks in the DGRP data, we removed one of the overlapping peaks from each pair in order to perform a paired test. We applied this procedure eight times to account for every possible combination of discarded peaks, finding that in all cases, before application of Z' or Z'', H_2/H_1 was greater in the DGRP data than in the DPGP data (P = 0.0473, averaged across the eight choices). After application of Z' and Z'', however, the comparison of DGRP and DPGP did not produce a significant difference (P = 0.1946 and P = 0.1781 for Z' and Z'', respectively, averaged across the eight choices). Thus, because normalization reduces the difference in H_2/H_1 values between corresponding peaks in the DGRP and DPGP data, the normalization suggests that differences in H_2/H_1 for corresponding peaks are attributable largely to the different values of H_{12} in the two data sets rather than to genuine differences in the biological signals that the two data sets provide.

Note that normalization can in principle change the rank order of peaks for a given data set, as a lower H_2/H_1 at a higher H_{12} can be shifted after normalization above a higher H_2/H_1 at a lower H_{12}. In our examples with the DGRP and DPGP data sets, however, relatively few reorderings of peaks took place upon normalization. We calculated a Spearman rank correlation coefficient to quantify the difference in rank order of Z and Z' values and Z and Z'' values for the overlapping peaks in the DGRP and DPGP data sets, and in all four calculations (DGRP Z to Z', DGRP Z to Z'', DPGP Z to Z', DPGP Z to Z''), the correlation coefficient exceeded 0.999.

Discussion

Statistical methods for detecting selective sweeps from genomic data have enabled the identification of cases of adaptation in multiple organisms. Many statistics have been developed to identify hard selective sweeps, and recent attention has now also focused on detecting soft sweeps (MESSER and NEHER, 2012; PETER *et al.*, 2012; FU and AKEY, 2013; MESSER and PETROV, 2013; VITTI *et al.*, 2013; FERRER-ADMETLLA *et al.*, 2014; JENSEN, 2014; WILSON *et al.*, 2014). GARUD *et al.* (2015) recently proposed the haplotype homozygosity statistics H_{12} and H_2/H_1 to discover both hard and soft selective sweeps and to differentiate whether top candidates for selection have signatures of hard or of soft sweeps. They applied their method to two *Drosophila* population-genomic data sets, DGRP and DPGP, recovering known cases of adaptation as well as finding new candidates.

In this paper, we have shown that the permissible range of H_2/H_1 values is dependent on their associated H_{12} values, and that therefore, the interpretation of H_2/H_1 in distinguishing hard and soft sweeps can be challenging when comparing H_2/H_1 values across loci with a broad distribution of H_{12} values. To facilitate interpretation of H_2/H_1 values measured in scans with a wide range of H_{12} values, we developed approximate and exact normalizations Z' and Z'' that can be applied to H_2/H_1. The application of the statistics Z' and Z'' to data has the greatest impact for H_2/H_1 values with high associated H_{12} values (>0.5).

We illustrated the use of the new bounds and normalizations using data from *Drosophila*. GARUD *et al.* (2015) compared the H_{12} peaks in the DGRP and DPGP data sets, finding that 13 DGRP peaks overlapped 16 DPGP peaks. However, the overlapping H_{12} peaks in the two data sets had significantly different H_2/H_1 values despite presumably reflecting the same selective events. In applying Z' and Z'' to the H_2/H_1 values observed at the highest, overlapping H_{12} peaks in the two data sets, we found that the comparison of distributions of H_2/H_1 values observed in the two scans did not produce a significant difference after normalization. Thus, the differences in distributions of H_{12} and H_2/H_1 across data sets might be attributable to differences in sample sizes and analysis window sizes in the two scans rather than to differences in biological signal. Indeed, the two data sets differed in a number of ways that could have generated higher H_{12} values on average for DPGP compared to DGRP. DPGP had a smaller sample size; in evaluating H_{12} from a finite sample of size $n \geq 2$, eq. 1 has a minimum of $(n + 2)/n^2$, which is greater for smaller n. H_{12} was also applied to DPGP in smaller analysis windows; decreasing the window size increases the probability of haplotype identity, thus increasing measures of homozygosity.

Our work on the relationship between H_{12} and H_2/H_1 parallels other studies (LONG and KITTLES, 2003; ROSENBERG *et al.*, 2003; HEDRICK, 2005; ROSENBERG and JAKOBSSON, 2008; VANLIERE and ROSENBERG, 2008; MARUKI *et al.*, 2012; REDDY and ROSENBERG, 2012; JAKOBSSON *et al.*, 2013; EDGE and ROSENBERG, 2014) in obtaining bounds on population-genetic statistics. A shared feature common to these studies is that in each study, unexpected or counterintuitive bounds are identified that are informative for sensible interpretation. As in some of these studies, however, our calculations consider an unspecified number of haplotypes K. If we instead required that K be specified as a finite constant, it would not be possible to reach the lower bound of 0 on H_2/H_1 because the lower bound is obtained from a limiting scenario with large numbers of low-frequency alleles. The difference in bounds between arbitrary-K and finite-K cases can for some statistics be nontrivial, especially for small K (Reddy and Rosenberg, 2012); for future work, it will be of interest to determine the magnitude of the effect on the H_2/H_1 bounds of fixing the value of K.

The proposed normalizations, Z' and Z'', offer an improvement in the interpretation of the H_{12} and H_2/H_1 statistics proposed by Garud *et al.* (2015). Further simulation-based investigation of the influence on H_{12} and H_2/H_1 of such variables as haplotype window sizes and sample sizes will be important for continuing to clarify the behavior of the statistics in models of selective sweeps. Nevertheless, as shown in our *Drosophila* example, the normalization of H_2/H_1 in data sets of varying sample sizes and SNP densities can help with the interpretation of selection scans, especially as data for testing population-genomic hypotheses become increasingly available in a variety of organisms.

Acknowledgments

We thank Doc Edge, Arbel Harpak, Rajiv McCoy, Pleuni Pennings, Dmitri Petrov, and Ben Wilson for helpful comments. Support was provided by NIH grant R01 HG005855.

References

Aminetzach YT, Macpherson JM, Petrov DA. Pesticide resistance via transposition-mediated adaptive gene truncation in Drosophila. Science. 2005; 309:764–767. [PubMed: 16051794]

Begun DJ, Aquadro CF. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature. 1992; 356:519–520. [PubMed: 1560824]

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics. 1995; 140:783–796. [PubMed: 7498754]

Catania F, Kauer MO, Daborn PJ, Yen JL, ffrench-Constant RH, et al. World-wide survey of an Accord insertion and its association with DDT resistance in Drosophila melanogaster. Molecular Ecology. 2004; 13:2491–2504. [PubMed: 15245421]

Comeron JM, Ratnappan R, Bailin S. The many landscapes of recombination in Drosophila melanogaster. PLoS Genetics. 2012; 8:e1002905. [PubMed: 23071443]

Cutter AD, Payseur BA. Genomic signatures of selection at linked sites: unifying the disparity among species. Nature Reviews Genetics. 2013; 14:262–274.

Daborn P, Boundy S, Yen J, Pittendrigh B, ffrench-Constant R. DDT resistance in Drosophila correlates with Cyp6g1 over-expression and confers cross-resistance to the neonicotinoid imidacloprid. Molecular Genetics and Genomics. 2001; 266:556–563. [PubMed: 11810226]

Depaulis F, Veuille M. Neutrality tests based on the distribution of haplotypes under an infinite-site model. Molecular Biology and Evolution. 1998; 15:1788–1790. [PubMed: 9917213]

Edge MD, Rosenberg NA. Upper bounds on FST in terms of the frequency of the most frequent allele and total homozygosity: the case of a specified number of alleles. Theoretical Population Biology. 2014; 97:20–34. [PubMed: 25132646]

Fay JC, Wu C-I. Hitchhiking under positive Darwinian selection. Genetics. 2000; 155:1405–1413. [PubMed: 10880498]

Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. On detecting incomplete soft or hard selective sweeps using haplotype structure. Molecular Biology and Evolution. 2014; 31:1275–1291. [PubMed: 24554778]

Fu W, Akey J. Selection and adaptation in the human genome. Annual Review of Genomics and Human Genetics. 2013; 14:467–489.

Garud NR, Messer PW, Buzbas EO, Petrov DA. Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. PLoS Genetics. 2015; 11:e1005004. [PubMed: 25706129]

Hedrick PW. A standardized genetic differentiation measure. Evolution. 2005; 59:1633–1638. [PubMed: 16329237]

Hermisson J, Pennings PS. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics. 2005; 169:2335–2352. [PubMed: 15716498]

Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ. Evidence for positive selection in the superoxide dismutase (sod) region of Drosophila melanogaster. Genetics. 1994; 136:1329–1340. [PubMed: 8013910]

Innan H, Kim Y. Pattern of polymorphism after strong artificial selection in a domestication event. Proceedings of the National Academy of Sciences of the United States of America. 2004; 101:10667–10672. [PubMed: 15249682]

Jakobsson M, Edge MD, Rosenberg NA. The relationship between FST and the frequency of the most frequent allele. Genetics. 2013; 193:515–528. [PubMed: 23172852]

Jensen JD. On the unfounded enthusiasm for soft selective sweeps. Nature Communications. 2014; 5:5281.

Jost L. GST and its relatives do not measure differentiation. Molecular Ecology. 2008; 17:4015–4026. [PubMed: 19238703]

Kaplan NL, Hudson RR, Langley CH. The “hitchhiking effect” revisited. Genetics. 1989; 123:887–899. [PubMed: 2612899]

Karasov T, Messer PW, Petrov DA. Evidence that adaptation in Drosophila is not limited by mutation at single sites. PLoS Genetics. 2010; 6:e1000924. [PubMed: 20585551]

Kim Y, Stephan W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics. 2002; 160:765–777. [PubMed: 11861577]

Long JC, Kittles RA. Human genetic diversity and the nonexistence of biological races. Human Biology. 2003; 75:449–471. [PubMed: 14655871]

Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, et al. The Drosophila melanogaster genetic reference panel. Nature. 2012; 482:173–178. [PubMed: 22318601]

Magwire MM, Bayer F, Webster CL, Cao C, Jiggins FM. Successive increases in the resistance of Drosophila to viral infection through a transposon insertion followed by a duplication. PLoS Genetics. 2011; 7:e1002337. [PubMed: 22028673]

Maruki T, Kumar S, Kim Y. Purifying selection modulates the estimates of population differentiation and confounds genome-wide comparisons across single-nucleotide polymorphisms. Molecular Biology and Evolution. 2012; 29:3617–3623. [PubMed: 22826460]

Maynard Smith J, Haigh J. The hitch-hiking effect of a favourable gene. Genetical Research. 1974; 23:23–35. [PubMed: 4407212]

Menozzi P, Shi MA, Lougarre A, Tang ZH, Fournier D. Mutations of acetylcholinesterase which confer insecticide resistance in Drosophila melanogaster populations. BMC Evolutionary Biology. 2004; 4:4. [PubMed: 15018651]

Messer PW, Neher RA. Estimating the strength of selective sweeps from deep population diversity data. Genetics. 2012; 191:593–605. [PubMed: 22491190]

Messer PW, Petrov DA. Population genomics of rapid adaptation by soft selective sweeps. Trends in Ecology and Evolution. 2013; 28:659–669. [PubMed: 24075201]

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, et al. Genomic scans for selective sweeps using SNP data. Genome Research. 2005; 15:1566–1575. [PubMed: 16251466]

Orr HA, Betancourt AJ. Haldane's sieve and adaptation from the standing genetic variation. Genetics. 2001; 157:875–884. [PubMed: 11157004]

Pennings PS, Hermisson J. Soft sweeps II—molecular population genetics of adaptation from recurrent mutation or migration. Molecular Biology and Evolution. 2006a; 23:1076–1084. [PubMed: 16520336]

Pennings PS, Hermisson J. Soft sweeps III: the signature of positive selection from recurrent mutation. PLoS Genetics. 2006b; 2:e186. [PubMed: 17173482]

Peter BM, Huerta-Sanchez E, Nielsen R. Distinguishing between selective sweeps from standing variation and from a de novo mutation. PLoS Genetics. 2012; 8:e1003011. [PubMed: 23071458]

Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Current Biology. 2010; 20:R208–R215. [PubMed: 20178769]

Przeworski M, Coop G, Wall JD. The signature of positive selection on standing genetic variation. Evolution. 2005; 59:2312–2323. [PubMed: 16396172]

Reddy SB, Rosenberg NA. Refining the relationship between homozygosity and the frequency of the most frequent allele. Journal of Mathematical Biology. 2012; 64:87–108. [PubMed: 21305294]

Rosenberg NA, Jakobsson M. The relationship between homozygosity and the frequency of the most frequent allele. Genetics. 2008; 179:2027–2036. [PubMed: 18689892]

Rosenberg NA, Li LM, Ward R, Pritchard JK. Informativeness of genetic markers for inference of ancestry. American Journal of Human Genetics. 2003; 73:1402–1422. [PubMed: 14631557]

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002; 419:832–837. [PubMed: 12397357]

Schmidt JM, Good RT, Appleton B, Sherrard J, Raymant GC, et al. Copy number variation and transposable elements feature in recent, ongoing adaptation at the Cyp6g1 locus. PLoS Genetics. 2010; 6:e1000998. [PubMed: 20585622]

Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989; 123:585–595. [PubMed: 2513255]

Teshima KM, Coop G, Przeworski M. How reliable are empirical genomic scans for selective sweeps? Genome Research. 2006; 16:702–712. [PubMed: 16687733]

VanLiere JM, Rosenberg NA. Mathematical properties of the r^2 measure of linkage disequilibrium. Theoretical Population Biology. 2008; 74:130–137. [PubMed: 18572214]

Vitti JJ, Grossman SR, Sabeti PC. Detecting natural selection in genomic data. Annual Review of Genetics. 2013; 47:97–120.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biology. 2006; 4:e72. [PubMed: 16494531]

Wilson BA, Petrov DA, Messer PW. Soft selective sweeps in complex demographic scenarios. Genetics. 2014; 198:669–684. [PubMed: 25060100]

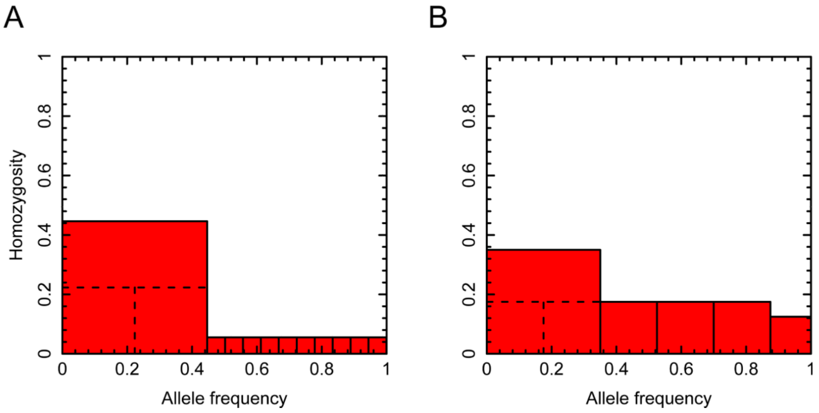

Figure 1.

A geometric illustration of the argument for finding the upper bound on H_2/H_1 as a function of H_{12}. In both panels, the unit interval (x-axis) is partitioned into components representing allele frequencies. H_{12} is represented by the sum of the areas of the red shaded regions, each indicating a squared frequency; the largest red square indicates $(p_1 + p_2)^2$ or M^2. (A) Step 1: for fixed H_{12} and fixed M, H_2/H_1 is maximal when $p_2 = p_1$. The maximal H_2/H_1 requires p_1 to be as small as possible, but $p_1 \geq p_2$ by definition; at the maximal H_2/H_1, p_1 and p_2 are equal. (B) Step 2: allowing M to vary while keeping H_{12} fixed, H_2/H_1 is maximal when M is as small as possible. At the maximum for H_2/H_1, M is reduced to the point where p_1 and as many subsequent alleles as possible have identical frequency, and at most one remaining allele of smaller frequency completes the unit interval. In both panels, H_{12}=0.23. Part A uses $\left(10+3\sqrt{170}\right)/110 \approx 0.4465$ for M and $\left(100-3\sqrt{170}\right)/1100 \approx 0.0553$ for each of 10 additional alleles. The dashed lines illustrate the choice of $p_2 = p_1 = M/2$. Part B achieves the maximum of $H_2/H_1 = 221/270 \approx 0.8185$ (eq. 12), with $M = 0.35$.

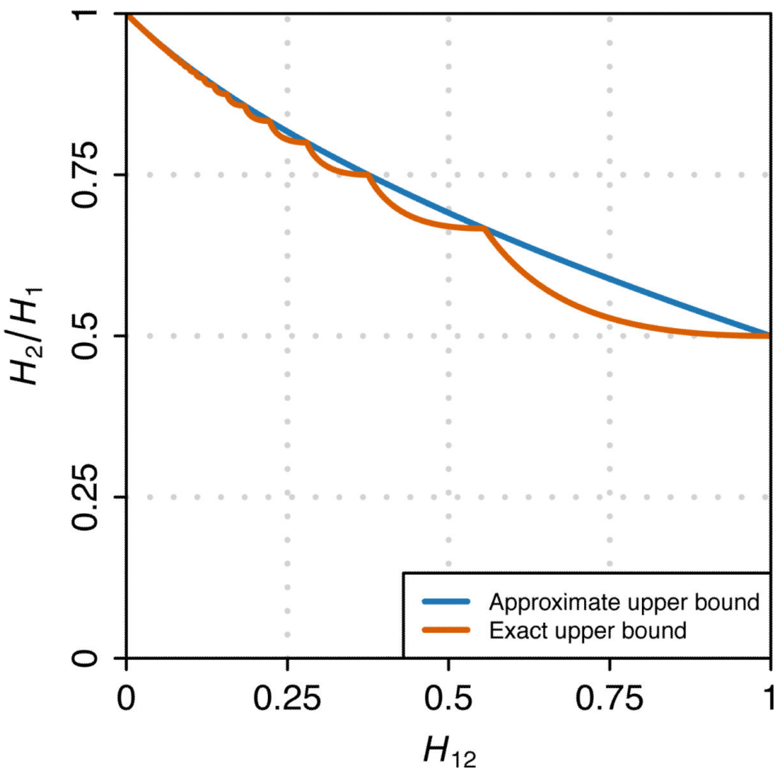

Figure 2.

The upper bound on H_2/H_1 as a function of H_{12}. The exact upper bound is given by eq. 12, and the approximate upper bound is given by eq. 15.

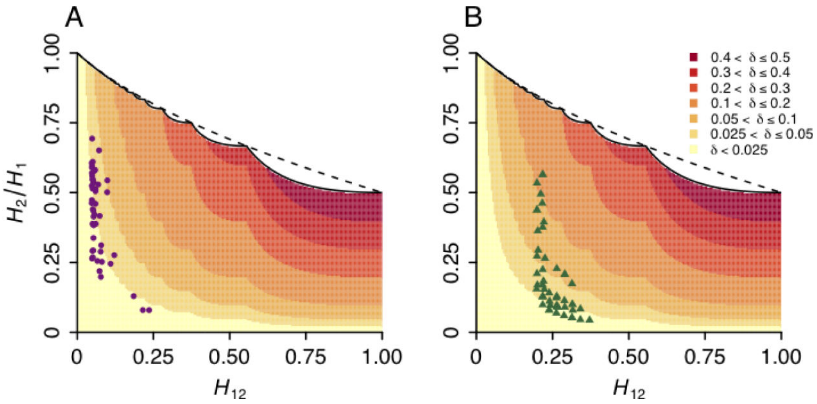

Figure 3.

The effect of the application of Z' on H_2/H_1 values in data from *Drosophila*. The shaded regions show the change δ in H_2/H_1 values after applying the normalization, where $\delta = Z' - Z$. Overlaid are points representing the top 50 windows for H_{12} in *Drosophila melanogaster* genome scans. (A) *Drosophila* Genetic Reference Panel (DGRP) data. (B) *Drosophila* Population Genomic Panel (DPGP) data. The solid line shows the exact upper bound on H_2/H_1 (eq. 12), and the dashed line shows the approximate upper bound (eq. 15).

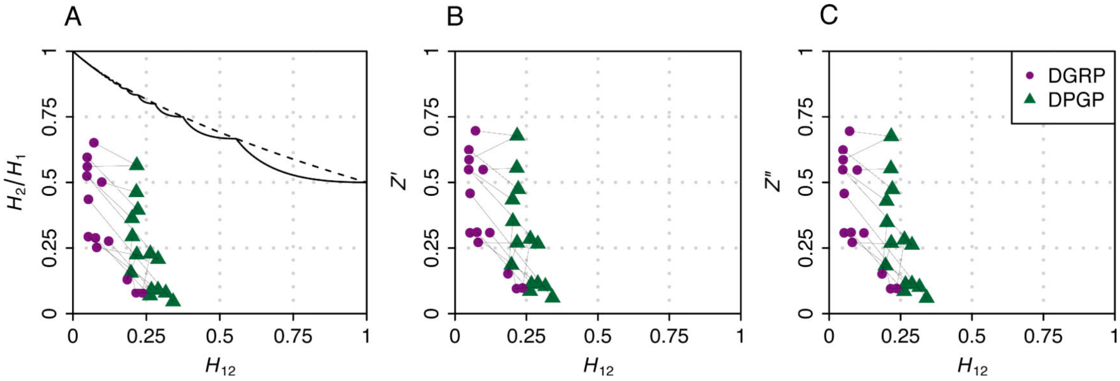

Figure 4.

H_{12} and H_2/H_1 values calculated in overlapping peaks in the DGRP and DPGP data sets before normalization and after the application of Z' and Z''. Corresponding points for the DGRP and DPGP data sets are connected by lines. Note that because the 16 DPGP peaks overlap 13 DGRP peaks, three DGRP points are each connected to a pair of DPGP points. Also, two pairs of DPGP points with different chromosomal locations have the same (H_{12}, H_2/H_1) coordinates. (A) Unnormalized H_2/H_1 values. Overlaid are the exact upper bound (solid) and the approximate upper bound (dashed) as given by eq. 12 and eq. 15. (B) Values of Z' (eq. 18). (C) Values of Z'' (eq. 19).