Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites

Hideki Innan,∗ Kangyu Zhang,† Paul Marjoram,‡ Simon Tavare,§ and Noah A. Rosenberg¶

Manuscript for Genetics, December 3, 2004

Several tests of neutral evolution employ the observed number of segregating sites and properties of the haplotype frequency distribution as summary statistics, and use simulations to obtain rejection probabilities. Here we develop a “haplotype configuration test” of neutrality (HCT) based on the full haplotype frequency distribution. To enable exact computation of rejection probabilities for small samples, we derive a recursion under the standard coalescent model for the joint distribution of the haplotype frequencies and the number of segregating sites. For larger samples, we consider simulation-based approaches. The utility of the HCT is demonstrated in simulations of alternative models and in application to data from Drosophila melanogaster.

Selectively neutral models of within-species evolution consist of a model that describes the genealogy of sampled DNA sequences, and a model for the stochastic process of mutation along the branches of the genealogy. Typical neutral models use the coalescent process (Nordborg, 2001, for example) to describe the genealogy, and the infinitely-many-sites model (Watterson, 1975) for the mutation process. Many theoretical predictions have been made under the standard neutral model, in which the particular coalescent model chosen is the one with constant population size.

As an alternative to computationally intensive comparisons of likelihoods of DNA sequence data under null and alternative models (Griffiths and Tavaré, 1994; Kuhner et al., 1998; Thomson et al., 2000), summaries of variation in a sample of sequences are often used in testing neutral models (Kreitman, 2000; Nielsen, 2001; Ford, 2002). Test statistics computed from the data are compared to theory-based predictions. If such predictions are unavailable or intractable, hypothesis testing is performed using simulations of the appropriate model.

Several neutrality tests use summary statistics based on the frequency distribution of haplotypes in a sample of DNA sequences from a particular region of a genome (Table 1). These “haplotype tests” are sometimes used to detect positive selection on particular haplotypes (Hudson et al., 1994). Alternatively, because haplotype frequency distributions may differ greatly across demographic scenarios (Donnelly, 1996; Nei et al., 1975), haplotype tests can also help to identify deviations from the demographic assumptions of the standard neutral model.

Some of the first haplotype tests, such as the Ewens-Watterson homozygosity test, were based on the Ewens (1972) sampling theory for the infinitely-many-alleles mutation model. Because “allele” in this model and “haplotype” in the infinitely-many-sites model have the same meaning, the Ewens (1972) theory provides the conditional distribution of the haplotype frequency vector C given the sample size n and the number of distinct haplotypes K, under the standard neutral model (Tavaré and Ewens, 1998, eq. 6). In the Ewens-Watterson test, haplotype homozygosity (F) is estimated for a sample of DNA sequences at a non-recombining locus as the sum of the squares of observed haplotype relative frequencies (Watterson, 1977, 1978). Given n, the value of F is compared to the known null distribution, and the neutral model is rejected if F is unusually high or low. A subsequent “exact test” was based on whether C itself was unlikely given n and K (Slatkin, 1994, 1996).

Other, more recently devised, haplotype-based tests reject the null hypothesis when the haplotype frequency distribution is unlikely under the same neutral model, given the number of mutations, or segregating sites (S),
TABLE 1: Haplotype-based tests of the neutral coalescent.

<table>
<thead>
<tr>
<th>Test</th>
<th>Quantity on which test is based</th>
<th>Reference for theoretical predictions about quantity</th>
<th>Reference for test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact test conditional on ( K )</td>
<td>( P(C</td>
<td>K,n) )</td>
<td>Ewens (1972)</td>
</tr>
<tr>
<td>Hudson et al. haplotype test (HHT)</td>
<td>( P(M</td>
<td>S,n,\theta) )</td>
<td>This article</td>
</tr>
<tr>
<td>Haplotype number test (HNT)</td>
<td>( P(K</td>
<td>S,n,\theta) )</td>
<td>Griffiths and Tavaré (1996), this article</td>
</tr>
<tr>
<td>Haplotype diversity test (HDT)</td>
<td>( P(H</td>
<td>S,n,\theta) )</td>
<td>This article</td>
</tr>
<tr>
<td>Haplotype configuration test (HCT)</td>
<td>( P(C</td>
<td>S,n,\theta) )</td>
<td>This article</td>
</tr>
</tbody>
</table>

Additional haplotype tests are discussed by Fu (1996, 1997), Kreitman (2000), and Sabeti et al. (2002). Fu (1998) provides a related but not haplotype-based test that uses the configuration of segregating sites in the sample.

TABLE 2: Notation.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Quantity represented</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n )</td>
<td>sample size</td>
</tr>
<tr>
<td>( \theta )</td>
<td>mutation parameter ((\theta = 2N\mu))</td>
</tr>
<tr>
<td>( N )</td>
<td>haploid population size (or scaling constant for the neutral coalescent model (Nordborg, 2001))</td>
</tr>
<tr>
<td>( \mu )</td>
<td>DNA sequence length times mutation rate per base pair per generation</td>
</tr>
<tr>
<td>( \rho )</td>
<td>recombination parameter ((\rho = 2Nr))</td>
</tr>
<tr>
<td>( r )</td>
<td>DNA sequence length times recombination rate per base pair per generation</td>
</tr>
<tr>
<td>( l )</td>
<td>DNA sequence length measured in base pairs</td>
</tr>
<tr>
<td>( I )</td>
<td>number of segments into which a recombination graph subdivides a DNA sequence</td>
</tr>
<tr>
<td>( \beta )</td>
<td>exponential population growth parameter</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>migration parameter in the two-island model in which both populations each have size ( N/2 ) ((\gamma = 2Nm))</td>
</tr>
<tr>
<td>( m )</td>
<td>in the symmetric two-island model, the fraction of individuals per population who migrate each generation</td>
</tr>
<tr>
<td>( S )</td>
<td>number of segregating sites</td>
</tr>
<tr>
<td>( M )</td>
<td>absolute frequency of the most frequent haplotype (that is, the number of sampled copies of the haplotype)</td>
</tr>
<tr>
<td>( K )</td>
<td>number of haplotypes</td>
</tr>
<tr>
<td>( H )</td>
<td>haplotype heterozygosity (or haplotype diversity)</td>
</tr>
<tr>
<td>( F )</td>
<td>haplotype homozygosity ((F = 1-H))</td>
</tr>
<tr>
<td>( G )</td>
<td>a univariate haplotype summary statistic in general (for example ( M, K, ) or ( H ))</td>
</tr>
<tr>
<td>( W_n )</td>
<td>waiting time until ( n ) lineages coalesce to ( n-1 ) lineages</td>
</tr>
<tr>
<td>( L )</td>
<td>total length of a genealogy with ( n ) lineages</td>
</tr>
<tr>
<td>( C = (C_1, C_2, ... , C_n) )</td>
<td>haplotype frequency vector (or haplotype configuration)</td>
</tr>
<tr>
<td>( C_k )</td>
<td>number of distinct haplotypes with absolute frequency ( k )</td>
</tr>
<tr>
<td>( \mathbf{e}_k )</td>
<td>vector with ( k )th coordinate 1 and other coordinates 0</td>
</tr>
<tr>
<td>( \delta_{i,j} )</td>
<td>Kronecker’s delta ((1 \text{ if } i = j, 0 \text{ otherwise}))</td>
</tr>
</tbody>
</table>

Corresponding lowercase letters are used for particular instances of random variables.

observed in the data (these tests are also conditioned on a mutation parameter \( \theta \), which for now we assume to be known). The adoption of a genealogical perspective has contributed to this shift in emphasis; in neutral models, conditional on the total length \((L)\) of the genealogical tree that underlies the data, \( E[S|L] \) is proportional to \( L \) (Hudson, 1990, for example). Thus, conditioning on \( S \) loosely serves as a proxy for conditioning on tree length.

Three main test statistics have been used in haplotype-based tests conditioned on \( S \) (DePaulis and Veuille, 1998; Hudson et al., 1994): the absolute frequency of the most frequent haplotype \((M)\), the number of distinct haplotypes \((K)\), and the haplotype diversity \((H)\). For each test statistic, many approaches are possible for implementing the test. We term tests based on \( M, K, \) and \( H \) the Hudson et al. haplotype test (HHT), haplotype
number test (HNT), and haplotype diversity test (HDT), respectively, allowing each name to apply to the collection of implementations of the relevant test.

Here we develop a “haplotype configuration test” of neutrality (HCT) conditional on $S$. This test is analogous to the exact test of SLATKIN (1994, 1996), which does not take $S$ into account. The HCT tests if the haplotype frequency vector $C$ is an unlikely configuration, given $S$ (and the mutation parameter $\theta$). The haplotype configuration $C$ conveniently summarizes the pattern of variation among DNA sequences; it perhaps incorporates more information about the data than do $M$, $K$, and $H$, but its use is less cumbersome than full-likelihood inference based on the DNA sequences themselves.

First, we derive a recursion for the joint distribution $\mathbb{P}(C,S|n,\theta)$, and we show how it can be used to obtain the conditional distribution $\mathbb{P}(C|S,n,\theta)$. This recursively-calculated conditional distribution allows exact computation of rejection probabilities for the HCT for small values of $S$ and $n$. Because $M$, $K$, and $H$ are calculated from $C$, the recursion can also enable exact computation of rejection probabilities for the HHT, HNT, and HDT. For large $S$ and $n$, use of the recursion is slow, and we discuss simulation techniques for all four tests. We describe simulation-based extensions that allow for more complex null hypotheses, which may include such phenomena as migration and recombination. We also consider various ways for addressing the problem that $\theta$ might not be “known”; in particular we argue that the effect of $\theta$ on haplotype tests can be addressed by using prior information about $\theta$ from genomic surveys. Finally, we consider the power of the HCT, HHT, HNT, and HDT against alternative models, and we apply the tests to an example from Drosophila melanogaster. The mathematical notation used here is shown in Table 2.

**THEORY**

**Joint distribution of $C$ and $S$:** We abbreviate $\mathbb{P}(C=c,S=s|n,\theta)$ by $q(c,s|n,\theta)$. This provides the probability that for $n$ lineages and mutation parameter $\theta$, the haplotype configuration is $c=(c_1,c_2,...,c_n)$ and $s$ segregating sites are observed, where $c_k$ is the number of haplotypes with absolute frequency $k$. We also abbreviate $\mathbb{P}(C=c|S=s,n,\theta)$ by $q(c|s,n,\theta)$ and $\mathbb{P}(C=c|n,\theta)$ by $q(c|n,\theta)$. We wish to express $q(c,s|n,\theta)$ in terms of probabilities for states one event backwards in time, where an “event” is either a coalescence of two lineages or a mutation. Using the standard neutral coalescent model with population size $N$ (NORDBORG, 2001, for example) and an infinitely-many-sites mutation model with mutation parameter $\theta$, time is scaled in units of $N$ generations so that the waiting time until the most recent coalescence is exponentially distributed with rate $n(n-1)/2$, and so that for each lineage, the waiting time until a mutation is exponentially distributed with rate $\theta/2$. Because $n$ lineages are present and because mutations on different lineages occur independently of one another, the waiting time (backwards in time) until a mutation happens on any lineage is exponentially distributed with rate $n\theta/2$. The probability that the most recent event is a coalescence is $(n-1)/(\theta + n - 1)$, and the probability that it is a mutation is $\theta/(\theta + n - 1)$, regardless of the time of the event. Suppose the event is a coalescence (a branching, if viewed forward in time). Then for each of $k=1,2,...,n-1$, it is possible that a haplotype of absolute frequency $k+1$ was generated by branching from a haplotype that had frequency $k$ in the previous step. The probability that the event produces a haplotype of frequency $k+1$ is the fraction of lineages whose haplotypes had frequency $k$ in the previous step, or $k(c_k+1)/(n-1)$.

If the event is a mutation, with probability $c_1/n$ the mutation occurs along a lineage of absolute frequency one, leaving the haplotype configuration unchanged. For each of $k=2,3,...,n$, it is also possible that a lineage whose haplotype has frequency $k$ experiences the mutation. In this situation, which has probability $k(c_k+1)/n$, a haplotype of frequency $k$ is replaced by a singleton haplotype and a haplotype of frequency $k-1$.

These cases are combined to produce the recursion:

\[
q(c,s|n,\theta) = \frac{\theta}{\theta + n - 1} \left[ \frac{c_1}{n} q(c-s|1,\theta) \right] + \sum_{k=2}^{n} \frac{k(c_k+1)}{n} \left[ q(c-e_k-s+1,\theta) \right] + \sum_{k=1}^{n-1} \frac{k(c_k+1)}{n-1} \left[ q(c+e_k-s+1,\theta) \right].
\]

In (1), $e_k$ is the $n$-dimensional vector with $k$th coordinate equal to 1 and all other coordinates equal to 0. For convenience, we treat all vectors as having length $n$, appending extra zeros to the ends of vectors of smaller lengths. We also specify the following rules:

\[
q(e_1,1,\theta) = \delta_{s,0} \quad (2)
\]
\[
q(c,s|n,\theta) = 0 \quad \text{if} \quad s < 0 \quad (3)
\]
\[
\text{For any} \ k, \ q(c,s|n,\theta) = 0 \quad \text{if} \quad c_k < 0 \quad (4)
\]
\[
q(c,s|n,\theta) = 0 \quad \text{if} \quad \sum_{k=1}^{n} kc_k \neq n \quad (5)
\]
\[
q(c,s|n,\theta) = 0 \quad \text{if} \quad \sum_{k=1}^{n} c_k > s + 1 \quad (6)
\]
\[
\text{For any} \ n, \ q(e_n,s|n,\theta) = 0 \quad \text{if} \quad s > 0. \quad (7)
\]

Induction using (1) with (2)-(7) can be used to derive various well-known results, as well as new expressions for
Expressions for the other three configurations with \( n = 4 \) are rather unwieldy and are not shown.

\( n = 3 \) and \( n = 4 \) (Table 3). For larger \( n \), the expression becomes unwieldy and numerical evaluation of (1) at particular \( \theta \) values is preferable (Table 4).

Summing (1) from \( s = 0 \) to \( \infty \) and rearranging terms produces the following recursion for \( q(c|n, \theta) \):

\[
[n(n-1) + (n-c_1)] q(c|n, \theta) = 
\theta \sum_{k=2}^{n} n(kc_k + 1) q(c - e_k - e_{k-1}, e_k|n, \theta)
+ n \sum_{k=1}^{n-1} k(c_k + 1) q(c + e_k - e_{k+1}|n-1, \theta). \tag{8}
\]

Conditions associated with (8) are obtained from sums of corresponding conditions for (1) from \( s = 0 \) to \( \infty \):

\[
q(e_1|1, \theta) = 1 \quad (9)
\]

For any \( k \), \( q(c|n, \theta) = 0 \) if \( c_k < 0 \) \quad (10)

\[
q(c|n, \theta) = 0 \quad \text{if} \quad \sum_{k=1}^{n} k c_k \neq n. \tag{11}
\]

It can be shown that (8) is equivalent to a recursion satisfied by the EWENS (1972) sampling formula, and used in its proof (Karlin and McGregor, 1972, eq. 9).

Thus, the solution to (8) with the initial condition (9) is the EWENS sampling formula (EWENS, 1972; Tavaré and EWENS, 1998, eq. 3):

\[
q(c|n, \theta) = \frac{n!}{\theta(n)} \prod_{k=1}^{n} \left( \frac{\theta}{k} \right)^{c_k} \frac{1}{c_k!}. \tag{12}
\]

where \( \theta(n) = \theta \cdot (\theta + 1) \cdots (\theta + n - 1) \), \( c_1, c_2, ..., c_n \) are non-negative integers, and \( \sum_{k=1}^{n} k c_k = n \). It is straightforward to verify that (12) indeed satisfies (8).

**Joint distribution of univariate haplotype summary statistics and S**: The statistics \( M, K, \) and \( H \) are functions of the frequency vector \( C \) as follows:

\[
M = \max\{k : C_k \neq 0\} \quad (13)
\]

\[
K = \sum_{k=1}^{n} C_k \quad (14)
\]

\[
H = 1 - \sum_{k=1}^{n} C_k \left( \frac{k}{n} \right)^2. \quad (15)
\]

For a given \( n \), only finitely many haplotype configurations are possible; thus, \( M, K, \) and \( H \) each have a finite range.
The probability that one of these statistics, \( G \), equals a value \( g \), is obtained by summing \( q(c,s|n,\theta) \) over the configurations that produce the value \( g \):

\[
P(G = g, S = s|n, \theta) = \sum_{c:G(c) = g} q(c,s|n,\theta).
\] (16)

**Conditioning on \( S \):** The conditional probability of \( C \), given \( S, n, \) and \( \theta \) is the quotient of the joint probability of \( C \) and \( S \) and the probability of observing \( S \) segregating sites:

\[
q(c|s,n,\theta) = \frac{q(c,s|n,\theta)}{P(S = s|n,\theta)}.
\] (17)

The denominator is the sum of the numerator over all configurations, and equals (Tavaré, 1984, eq. 9.5):

\[
P(S = s|n, \theta) = \frac{n - 1}{\theta} \sum_{j=1}^{n-1} (-1)^{j-1} \binom{n-2}{j-1} \left( \frac{\theta}{\theta+j} \right)^{s+1}.
\] (18)

Similar equations to (17) apply for \( M, K, \) and \( H \); if \( G \) represents one of these statistics, then

\[
P(G = g|S = s, n, \theta) = \frac{P(G = g, S = s|n, \theta)}{P(S = s|n, \theta)}.
\] (19)

**STATISTICAL TESTS**

The exact computation of \( q(c|s,n,\theta) \) suggests the following haplotype configuration test (HCT).

**Procedure 1:** Exact implementation of the haplotype configuration test.

1. Under the null hypothesis of neutrality, use (17) to compute the probabilities of all haplotype configurations, given \( s, n, \) and the known \( \theta \).

2. Sum the probabilities of all haplotype configurations whose probabilities under the null model are less than or equal to the probability of the observed configuration \( c \). That is, compute

\[
P = \sum_{\{b : q(b|s,n,\theta) \leq q(c|s,n,\theta)\}} q(b|s,n,\theta).
\] (20)

3. Reject the null hypothesis at level \( \alpha \) if \( P \leq \alpha \).

Two-tailed haplotype tests based on univariate summary statistics can be implemented in a similar manner (one-tailed versions of these tests are also possible).

**Procedure 2:** Exact implementation of (two-tailed) haplotype tests based on univariate summary statistics.

1. Under the null hypothesis of neutrality, use (19) to compute the probabilities of all possible values of the summary statistic \( G \) given \( s, n, \) and the known \( \theta \).

2. If \( g \) denotes the observed value of \( G \) in the data, reject the null hypothesis at level \( \alpha \) if \( P(G \leq g) \leq \alpha/2 \) or if \( P(G \geq g) \leq \alpha/2 \).

For the tests based on \( M, K, \) and \( H \), the null hypothesis is rejected if a particular aspect of the observed haplotype frequency distribution is a rare occurrence. Using the HCT, however, the null hypothesis is rejected when the observed haplotype frequency distribution itself is rare. Rare configurations will typically — but not always — have unlikely values for one or more of \( M, K, \) and \( H \).

For three choices of \( \theta \), Table 5 shows exact probabilities for \( C, M, K, \) and \( H \), given \( s = 10 \) and \( n = 7 \), obtained numerically from (17) and (19). The probabilities shown provide rejection probabilities for the four tests. In Table 5, certain haplotype configurations may be unlikely, even if their values of \( M, K, \) and \( H \) are centrally located with respect to null distributions of these statistics. For example, \( (0,2,1,0,0,0,0) \) is highly unlikely for \( s = 10 \) and \( \theta \approx 4.08 \): the HCT rejects the null hypothesis at \( P = 0.0285 \). However, the frequency of the most frequent haplotype (3), the number of haplotypes (3), and the haplotype diversity (0.6531), are all rather ordinary for the given parameter values.

It is also possible that from among the HHT, HNT, and HDT, one or more tests might reject the null hypothesis at smaller \( \alpha \) than does the haplotype configuration test. For example, in Table 5 with \( \theta \approx 4.08 \), \( (7,0,0,0,0,0,0) \) is unusual at the \( \alpha = 2 \times 0.0328 = 0.0656 \) level for the HHT, HNT, and HDT, but only at \( \alpha = 0.1113 \) using the HCT.
**P-values for small samples:** Because C, M, K, and H can take on only finitely many values for a given n, there is some probability that (for example) \( \mathbb{P}(G < g) \leq \alpha/2 \) but \( \mathbb{P}(G \leq g) > \alpha/2 \). In such cases, Procedures 1 and 2 do not reject the null hypothesis and thus are slightly conservative. This situation often arises with small samples, for which not many numbers can be possible values of the test statistics. It occurs most often for \( M \) and \( K \), each of which has only \( n \) possible values. A small-sample correction for the HCT is to append to Step 3 of Procedure 1: If \( P > \alpha \) but \( P - Q < \alpha \), where

\[
Q = \sum_{\{b \mid q(b|s,n,\theta) = q(c|s,n,\theta)\}} q(b|s,n,\theta), \tag{21}
\]

reject the null at level \( \alpha \) with probability \( (\alpha - P + Q)/Q \).

For the other tests, we can append to Step 2 in Procedure 2: If \( \mathbb{P}(G < g) > \alpha/2 \) but \( \mathbb{P}(G < g) < \alpha/2 \), reject the null hypothesis at level \( \alpha \) with probability \( (\alpha/2 - \mathbb{P}(G < g))/\mathbb{P}(G = g) \); if \( \mathbb{P}(G \geq g) > \alpha/2 \) but \( \mathbb{P}(G > g) < \alpha/2 \), reject the null hypothesis at level \( \alpha \) with probability \( (\alpha/2 - \mathbb{P}(G > g))/\mathbb{P}(G = g) \).

Corrections for one-tailed versions of the HHT, HNT, and HDT are analogous. The conservative Procedures 1 and 2 are suitable for data analysis; the small-sample correction is most useful when it is important for a rejection region to have fixed size, as in evaluations of the power to reject the null hypothesis under alternative models.

**SIMULATION-BASED IMPLEMENTATIONS**

Thus far, we have assumed that the sample size and number of segregating sites are small enough that numerical iteration of the recursion does not hold. We have also assumed a simple demographic model without recombination and that \( \theta \) is known exactly. We now discuss implementations of the HCT and the other tests when these ideal conditions do not hold. First, we consider simulation methods, demographic models, and recombination, continuing to assume that \( \theta \) is known; we then discuss ways in which uncertainty in \( \theta \) might be incorporated.

**Methods for simulation:** Earlier articles on haplotype tests have described various simulation-based implementations. These algorithms will be generally applicable to the HCT as well. Although exact computation from (17) and (19) is appropriate for small \( s \) and \( n \), simulation is necessary as \( s \) and \( n \) increase.

The procedures that simulate from the correct conditional distribution \( \mathbb{P}(C|S,n,\theta) \) might be classified as Markov chain Monte Carlo methods (Markovtsova et al., 2000, 2001), importance sampling methods (Dépaulis et al., 2001), and acceptance-rejection algorithms (Tavaré et al., 1997; Wall and Hudson, 2001). A simple and often highly efficient approach is the following version of Algorithm 1 of Tavaré et al. (1997).

**Procedure 3: Acceptance-rejection algorithm for generating samples from \( \mathbb{P}(C|S,n,\theta) \).**

1. Simulate the coalescence times \( W_n, W_{n-1}, \ldots, W_2 \) as independent exponentially distributed random variables, with \( W_j \sim \text{exp}(j-1)/2 \).
2. Compute the total branch length of the resulting genealogical tree, \( L = \sum_{j=2}^{n-1} W_j \).
3. Accept the simulated collection of \( W_j \) values with probability \( u \), where

\[
u = \frac{e^{-L\theta/2}(L\theta/2)^S}{e^{-S\theta}S^S}. \tag{22}\]

Otherwise, discard the \( W_j \) and return to the initial step.
4. Simulate the branching structure of the genealogy by randomly joining lineages until one lineage remains, and associate the \( W_j \) with corresponding branching events.
5. Independently place \( S \) mutations uniformly on the genealogy.
6. Record the haplotype configuration \( C \) (and the quantities \( M, K, \) and \( H \)).

Procedure 3 is then repeated until a pre-specified number of genealogies have been accepted. The empirical distribution of \( C \) is then used in Procedure 1 to decide whether or not to reject the null hypothesis. To implement the HHT, HNT, and HDT, the empirical distributions of \( M, K, \) and \( H \) are used in the final step of Procedure 2.

The efficiency of Procedure 3 derives from the fact that it accepts or rejects the simulation before placing \( S \) mutations on the branching diagram. It is less efficient to place mutations on the genealogy with a Poisson process with mean \( L\theta/2 \) and only afterwards accept trees that have accumulated exactly \( S \) mutations. The denominator of (22) is the maximum of the numerator over all values of \( L \), ensuring that acceptance will occur reasonably often, except if \( S \) is much larger or much smaller than is suggested by the value of \( \theta \) (Tavaré et al., 1997).

It is noteworthy that even with the efficient algorithm of Procedure 3, for large \( n \), due to the large number of possible configurations, the HCT has computational limitations not shared by the other tests. For a given sample size \( n \), the number of possible values of the test statistics for the HHT or HNT is only \( n \). For the HDT, up to a linear transformation, the set of possible values of the statistic is equal to the set of numbers that can equal the sum of the squares of the elements of a partition of \( n \) into positive integers (On-Line Encyclopedia of Integer Sequences, entry A069999). The configurations that produce the smallest and largest sums of squares, which equal \( n \) and \( n^2 \), respectively, are \((n,0,...,0)\) and \((0,0,...,0,1)\). Because the parity of the sum of squares
in any partition of \( n \) must be the same as that of \( n \), \((n^2 - n + 2)/2\) provides an upper bound on the number of possible values of the HDT statistic. Even for sample sizes that are presently considered large, using a large number of accepted simulated genealogies (exceeding the number of possible values of the test statistic by a factor of \( \geq 100 \)), it is feasible to approximate the probabilities of all possible values of \( M, K, \) and \( H \).

Unlike the other tests, however, the HCT has the form of an exact test; such tests are characterized by enumeration of the probabilities of all possible data sets under the model, and summation of the probabilities of all data sets that are at most as probable as the observed data set (MEHTA and PATEL, 1998). For these tests, the number of possible data sets increases very rapidly with some property of the data (such as the number of alleles at the locus, in exact tests of Hardy-Weinberg equilibrium). For the HCT, the number of haplotype configurations, equivalent to the number of unordered partitions of \( n \) into positive integers, \( p(n) \), grows quickly with \( n \) (ABRAMOWITZ and STEGUN, 1965, p. 836). While \( n = 10 \) has only 42 partitions, \( n = 50 \) has 204,226, and there are \( \approx 4 \times 10^{12} \) for \( n = 200 \). Because enumeration of the probabilities of all configurations is impossible for large samples, the HCT is limited to sample sizes \( n \) for which it is feasible to generate \( \geq 100p(n) \) samples from \( P(C|S,n,\theta) \). As Markov chain Monte Carlo algorithms have been developed for other exact tests in genetics to avoid the enumeration problem (GUO and THOMPSON, 1992; RAYMOND and ROUSSEt, 1995), it is conceivable that such an algorithm might be developed for this test as well.

**Demographic models:** The approach in Procedure 3 is versatile, in that the null model need not be the standard constant-size coalescent. A more complex demographic model can be accommodated by substituting its distributions for the waiting times and the branching structure in Steps 1 and 4, respectively.

For example, in place of a population with constant size \( N \), consider an exponentially growing population of present population size \( N \) with growth parameter \( \beta \). At \( t \) time units of \( N \) generations in the past, population size was \( N \exp[-\beta t] \). To use this model as the null, after Step 1, for each \( j \), replace \( W_j \) by \( f(W_j) \), where (SLATKIN and HUDSON, 1991; NORDBORG, 2001, eq. 8)

\[
 f(W_n) = \frac{\ln(1 + \beta W_n)}{\beta},
\]

(23)

and for \( j \in \{2, 3, \ldots, n - 1\} \),

\[
 f(W_j) = \frac{\ln \left(1 + \beta \sum_{k=j}^{n} W_k\right) - \ln \left(1 + \beta \sum_{k=j+1}^{n} W_k\right)}{\beta}.
\]

(24)

For each \( j \), \( f(W_j) \) gives the time to coalescence of \( j \) lineages to \( j - 1 \), measured in units of \( N \) generations.

Because variable population size does not change the branching structure of the genealogy, the rest of Procedure 3 is identical under exponential population growth.

With a population structure model, such as island migration, as the null, it is simpler to simulate the appropriate waiting times and branching structure concurrently in place of Step 1 (HUDSON, 1990), omitting Step 4.

**Recombination:** Using the ancestral recombination graph (GRIFFITHS and MARJoram, 1996; HUDSON, 1983; NORDBORG, 2001), Procedure 3 can be extended to allow recombination in the null model. Suppose the recombination parameter for a DNA sequence region is \( \rho = 2Nr \), where \( l \) is the number of base pairs in the region and \( r \) is \( l \) times the recombination rate per base per generation. Simulation of the ancestral recombination graph for \( n \) lineages consists in repeatedly simulating an exponentially distributed random variable for the time of the next coalescence (appropriately transformed if the model includes population growth), and another exponential random variable for the time of the next recombination. The smaller of the two times gives the type of the next event, which is then allowed to occur, the larger time is discarded, and uniform random variables are simulated to decide which lineages participate in the event. From the graph, a genealogical tree is produced at each base pair, so that nearby base pairs are more likely to have equivalent trees than are distant base pairs (ROSENBERG and NORDBORG, 2002, figure 3, for example). The graph splits the sequence into \( I \) segments with lengths \( l_1, \ldots, l_I \), such that all base pairs within a segment have equivalent genealogies, and such that recombination events occur only at segment boundaries. Denote the total branch length of the genealogy of segment \( i \) by \( L_i \).

**Procedure 4:** Acceptance-rejection algorithm for generating samples with recombination from \( P(C|S,n,\theta,\rho) \).

1. Simulate an ancestral recombination graph for \( n \) lineages with recombination rate \( \rho \), until all parts of the DNA sequence reach most recent common ancestors.

2. Compute the total branch length of the genealogical tree at an "average" site, \( L = \sum_{i=1}^{I} L_i / l \).

3. Accept the simulated graph with probability \( u \), where

\[
 u = \frac{e^{-L\theta/2}(L\theta/2)^S}{e^{-S^2/2}}.
\]

(25)

Otherwise, discard the graph and return to the initial step.

4. Independently place \( S \) mutations uniformly on the genealogy.

5. Record the haplotype configuration \( C \) (and the quantities \( M, K, \) and \( H \)).
This procedure enables haplotype tests to be performed conditional on a known recombination parameter as well as on the known mutation parameter. In the computer program we have implemented, which proceeds more slowly than the algorithm in Procedure 4, placement of mutations occurs concurrently with the simulation of the graph in Step 1. In place of Step 3, the graph is accepted if $S$ mutations are placed, and Steps 2 and 4 are omitted.

**Treatment of $\theta$:** Because $\theta$ does not affect the conditional distribution of $C$ given the sample size and the number of haplotypes, its value is not of concern in neutrality tests based on the Ewens (1972) sampling theory. However, $\theta$ does affect $P(C|S, n, \theta)$ and analogous distributions for $M, K$, and $H$. As an example, in Table 5, for $\theta = 1$, the most common and rarest configurations are $(2,1,0,0,0,0)$ and $(7,0,0,0,0,0)$, respectively, while $(5,1,0,0,0,0,0)$ is most common and $(0,0,1,0,0,0,0)$ is rarest for $\theta = 10$. Thus, as has been reported for other neutrality tests (Fu, 1996, 1998; Markovtsova et al., 2001; Wall and Hudson, 2001), it might be expected that the value of $\theta$ used can affect rejection probabilities for the HCT, as well as for the HHT, HNT, and HDT.

Table 5 indicates that nominal and actual rejection probabilities for the HCT can differ substantially if the value of $\theta$ used is a poor estimate. For example, suppose that the true $\theta$ equals 1 and that the configuration $(1,0,2,0,0,0,0)$ is observed with $s = 10$. The actual HCT rejection probability is $P = 0.1748$. If no prior knowledge about $\theta$ was available, a sensible procedure would be to base $P$-values on a value of $\theta$ estimated using an estimator such as that of Watterson (1975),

$$\hat{\theta}_W = S / \left( \sum_{i=1}^{n} 1/i \right).$$

Equation (26) gives $\hat{\theta}_W \approx 4.08$; had this estimate been used, it would have been concluded that the observed configuration is significantly unlikely at $\alpha = 0.05$ ($P \approx 0.0477$; see the middle part of Table 5).

This variation in $P$-values as $\theta$ fluctuates suggests that for tests that depend on $\theta$, the standard method for obtaining rejection probabilities when they vary with the unknown value of a nuisance parameter — using the maximal rejection probability across possible values of the parameter (Berger and Boos, 1994) — is overly conservative. Indeed, in the example of $s = 10$ and $n = 7$, for each of the 14 haplotype configurations possible with $s > 0$, there is some value of $\theta$ for which the configuration is reasonably likely (Figure 1A). In this example, if the largest rejection probability across all values of $\theta$ were used, for the HCT, no configuration would lead to rejection at $\alpha = 0.05$, or even at $\alpha = 0.2$ (Figure 1B).

Because point estimates of $\theta$ generally have large variances (Felsenstein, 1992; Fu and Li, 1993), a strategy of maximizing the rejection probability only over a confidence set for $\theta$ (Berger and Boos, 1994), and not over the full range $(0, \infty)$, is also likely to be quite conservative. This is especially true as $P$-values can vary considerably in the vicinity of a point estimate (consider Figure 1B at $\log_{10} \hat{\theta}_W \approx 0.61$). Thus, an alternative method is to substitute a distribution of values for $\theta$ in a Bayesian procedure (Fu, 1998; Kelly, 1997; Depaulis et al., 2001).

In this scheme, a prior probability density, $f_{\text{prior}}(\theta)$, is chosen, and rejection probabilities are evaluated using

---

**FIGURE 1:** Haplotype configuration test rejection probabilities as functions of $\theta$, for $s = 10$ and $n = 7$. (A) Probabilities of haplotype configurations, obtained from (17). (B) $P$-values for the HCT, obtained from (20). For each value of $\theta$ at which two configurations have equal probability — that is, when the curves for the configurations cross in (A) — HCT $P$-values for the two configurations both experience discontinuities in $\theta$. In (B), data for a given configuration are shown with the same color and pattern as used to depict the configuration in (A).
the density that is proportional to \( P(C|S, n, \theta)f_{\text{prior}}(\theta) \). To implement this approach, we can add a Step 0 to Procedure 3: **Simulate a value of \( \theta \) from \( f_{\text{prior}}(\theta) \).** Under exponential growth or recombination, \( \beta \) and \( \rho \) can also be chosen from priors, so that evaluation of rejection probabilities is performed conditional on the prior rather than on a fixed growth or recombination rate. Using a uniform prior on \([0,20]\) for \( \theta \) (with no growth or recombination), we implemented Procedure 3 with \( s = 10 \) and \( n = 7 \) (Table 6). The distribution for the Bayesian approach is similar to that for \( \hat{\theta} = 4.08 \); for the example configuration \((1.0,2.0,0,0,0,0)\), the HCT rejection probability is 0.0403, close to the value obtained using the point estimate. Other priors, such as uniform distributions on \([0,100]\) or \([0,500]\), produce similar results (not shown).

An alternate approach to uncertainty in \( \theta \) is a simulation strategy that does not explicitly use \( \theta \). In this procedure (Hudson, 1993; Hudson et al., 1994), equivalent to assuming an implicit density, \( f_{\text{imp}}(\theta|S, n) \), and simulating from a density proportional to \( P(C|S, n, \theta)f_{\text{imp}}(\theta|S, n) \), the waiting times and branching structures of genealogies are simulated under the neutral coalescent. On each genealogy, \( S \) mutations are placed, uniformly and independently. The null distribution of \( C \) is taken as the empirical distribution of its values for the simulated trees. If the estimate \( \hat{\theta}_W \) is close to the true value \( \theta \), this procedure produces similar rejection probabilities to both the fixed-\( \theta \) approach that assumes \( \theta = \hat{\theta}_W \) and the Bayesian procedure in the previous paragraph. However, if \( \hat{\theta}_W \) and \( \theta \) differ substantially, the conditional distributions of test statistics by this kind of fixed-\( S \) simulation may also be quite different from their conditional distributions given \( S \) and \( \theta \) (Markovtsova et al., 2001; Wall and Hudson, 2001). This observation is supported also by our simulation with 10 mutations and \( n = 7 \) (Table 6). The probability distribution for fixed-\( S \) simulation is quite similar to the exact distribution shown in Table 5 with \( \theta = \hat{\theta}_W \), and the HCT rejection probability for the example configuration \((1.0,2.0,0,0,0,0)\) is 0.0520.

Tables 5 and 6 and Figure 1 show that the effect of \( \theta \) on haplotype tests is not negligible, and that erroneous conclusions might be reached if the observed number of segregating sites is not close to expectation. The problem is not fixed by using a distribution of values for \( \theta \) estimated from the same data that are to be analyzed using the tests, as such a distribution will likely produce results similar to those obtained with a point estimate. However, the use of diverse regions spread across a genome decreases the variance of an estimate of \( \theta \) dramatically (Innan et al., 2003). Thus, with genomic polymorphism data, \( \theta \) can be estimated at many loci independent of the region of interest, and the genomic estimate can be treated as the “known” value of \( \theta \). A collection of values spread around a genomic point estimate might also be used in a Bayesian procedure, though this approach will probably not differ greatly in outcome from use of only the point estimate.

The application of a genomically estimated \( \theta \) requires the assumption of a constant value of \( \theta \) across the genome. However, variables such as GC-content and recombination rate may lead to considerable variation in \( \theta \) (Begun and Aquadro, 1992, for example). In such cases, the “known” \( \theta \) could be estimated within classes of regions that have similar GC-content, levels of recombination, or values of other quantities that influence \( \theta \).

**POWER**

We investigated the power of the four haplotype tests to detect deviations from the standard coalescent model. For various choices of \( s \) and \( n \), we used Procedure 3 with \( \theta_W \) for \( \theta \) to obtain the distributions of \( C \), \( M \), \( K \), and \( H \) under the null model. For given choices of \( s \) and \( n \), the empirical null distributions of \( C \), \( M \), \( K \), and \( H \) were based on a set of \( 10^5 \) accepted simulations; for \( n \leq 30 \), this number was found sufficient to ensure that the distributions were accurately estimated. The rejection regions for the tests were defined using Procedures 1 and 2, employing simulated probabilities in place of the exact probabilities, and using the correction for small-sample size: haplotypes on the rejection region boundary for the chosen significance level were assigned an appropriate rejection probability in \((0.1)\), and all other haplotypes were placed either inside or outside the rejection region.

For each choice of \( s \) and \( n \) and each alternative model, the power to reject the null for significance level \( \alpha \) was equal to the fraction of \( 10^5 \) simulations of the alternative whose haplotypes lay in the \( \alpha \)-rejection region. Note that in contrast to the simulations of the null model, which were used to simultaneously estimate a large number of quantities, namely the probabilities of all possible haplotype frequencies or values of a test statistic, simulations of the alternative were used only to estimate what fraction of replicates lay in the rejection region. Thus, in comparison with the null model, to obtain repeatable results, the alternative model required fewer replicates.

Simulations of the alternative model require use of a value of \( \theta \). As discussed earlier, when the null hypothesis is true, a sensible choice is to use a value of \( \theta \) estimated by assuming that the null is true — for example, \( \hat{\theta}_W \). Thus, when the alternative hypothesis is true, a sensible choice of \( \theta \) on which to condition the simulations is a value that would have been estimated assuming that the alternative was true. For a general model in which the expected total branch length of a genealogical tree is \( E[L] \), the generalized expression analogous to \( \hat{\theta}_W \) is \( 2S/E[L] \).

In models with recombination, the recombination rate affects the variation across sites in branch lengths of ge-
FIGURE 2: Power of four tests to reject the standard neutral model at significance level $\alpha = 0.05$, for three values of $s$ (10, 30, 75) and three types of alternative models — from top to bottom, recombination, exponential population growth, and island migration. In each plot, $n = 30$ (in the migration model, the sample was separated into two populations each with sample size 15). Simulation of the null model used Procedure 3; simulation of the alternative model used Procedure 4 in the case of recombination, and appropriately modified versions of Procedure 3 in the cases of exponential population growth and island migration (see section on Demographic models).

Recombination: If no recombination takes place, the strong correlation of genotypes at neighboring sites makes it fairly likely that if two DNA sequences both contain a mutation at a particular site, they will also share the same haplotype. Thus, many individuals may have the same haplotype, and configurations with high $M$, low $K$, and low $H$ are common.

Recombination decouples neighboring sites, so that for large recombination rates, neighboring genotypes are nearly independent. Thus, two sequences with the same genotype at one site will have the same haplotype only if at each of the $s - 1$ remaining sites, they share the same genotype — an unlikely event whose chance of occurring is the product of $s - 1$ probabilities. Consequently, given $s$ and $n$, as recombination rate increases, $M$ decreases,

nealogical trees but not the expected total branch length of a randomly chosen site (PLUZHNIKOV and DONNELLY, 1996). Thus, the expected total branch length is the same as in the absence of recombination, and consequently, we used $\hat{\theta}_W$ for the simulations of the alternative model.

For exponential population growth and two-population symmetric island migration models, however, the branch lengths of genealogies differ from those of the standard model. Our simulations of the alternative model employed $2s/E[L]$ for $\theta$, where $E[L]$ was the mean total branch length in $10^5$ simulations of the model. Note that this choice makes the value of $\theta$ at which tests were performed dependent on the parameters of the alternative model. Once the value of $\theta$ was selected, an independent set of simulations was used for the estimation of power.
$K$ and $H$ increase, and $C$ tends towards configurations that contain many haplotypes of frequency 1, such as $se_1 + e_{n-s}$ for small $s$ and $ne_1$ for large $s$.

For relatively large recombination rates, all tests were able to reject the null model in favor of the recombination model, with power greater in the scenarios with smaller $s$ (Figure 2). For small $s$, power was comparable for the four tests, whereas for intermediate and large $s$, the HCT had poorer relative performance than the other tests. The HNT performed particularly well, in agreement with previous observations about the informativeness of $K$ about the recombination rate (SCHIERUP and HEIN, 2000; WALL, 2000; MYERS and GRIFFITHS, 2003).

**Exponential population growth:** Because coalescences are more probable in smaller populations, by increasing the population size in recent generations compared with that of ancient generations, exponential population growth makes it more likely that lineages will persist into the distant past without coalescing. Thus, growth increases lengths of terminal branches of coalescent trees in comparison with those of internal branches. Mutations on terminal branches create single-copy haplotypes, so that growth increases the frequencies of configurations with low $M$, high $K$, and high $H$. As in the recombination case, the configurations $se_1 + e_{n-s}$, and $ne_1$ become probable for small and large $s$, respectively.

In the simulations with exponential growth, the HCT had comparable power to the HNT for small and intermediate $s$ (Figure 2). The HHT and HDT performed rather poorly for small $s$, improving as $s$ increased. As in the recombination simulations, the HNT was the most generally powerful test. Although some exceptions were observed (for example, with $s = 10$), similar results were usually obtained when the null hypothesis included recombination (Figure 3).

**Island migration:** With a large migration rate, the behavior of the two-population island migration model approaches that of the standard neutral model (NORDborg and KRONE, 2002, for example). Thus, haplotype frequencies for large migration rates will be similar to those under the null model. As the migration rate decreases, however, lineages coalesce separately in the two populations, with no intervening migrations. The time until one of these lineages migrates to the other population is long, so that genealogies have two long internal branches. These branches contain most mutations, leading to configurations with high $M$, low $K$, and low $H$.

The simulations produced reasonable power with low migration rates, and low power with high migration rates (Figure 2). The HCT and HNT had comparable power for intermediate and large $s$. Similarly to the population growth simulations, the HHT and HDT performed poorly for small $s$, with similar results when recombination was included in the null hypothesis (Figure 3). As in both the recombination and the growth scenarios, the HNT was the most generally powerful test for small $s$.

FIGURE 3: Power of four tests to reject the standard neutral model at significance level $\alpha = 0.05$ for a null model including recombination ($\rho = 10$), and two types of alternative models — exponential population growth (top) and island migration (bottom). Simulation of the null model used Procedure 4; simulation of the alternative model used appropriately modified versions of Procedure 4. The simulations were otherwise the same as in Figure 2.
APPLICATION TO DATA

In a data set for the Sod locus in Drosophila melanogaster (Hudson et al., 1994), 55 segregating sites were observed in a sample of size 10, with \( C = (5, 0, 0, 0, 1, 0, 0, 0, 0, 0) \). Thus, \( M = 5, K = 6, \) and \( H = 0.7 \). Using (26), \( \hat{\theta}_E \approx 19.44 \). To demonstrate the application of the four haplotype tests, we used Procedures 3 and 4 (appropriately modified in the case of the application of the four haplotype tests, we used Probesites were observed in a sample of size 10, with \( 0 \), \( 0 \), \( 0 \), \( 0 \), \( 0 \), \( 0 \)).

As was observed by Hudson et al. (1994), \( M = 5 \) is highly unusual under the standard neutral model, as well as under models that include small amounts of population growth and recombination. In each case, the HHT and HDT reject the null hypothesis at very low \( P \)-values. The HCT is significant below the 5% level in all cases, while the two-tailed HNT is not significant at the 5% level for low levels of growth and recombination.

CONCLUSIONS

We have introduced the haplotype configuration test of neutrality, which rejects the null hypothesis if the configuration itself is unlikely given \( S, n, \) and \( \theta \). We have also developed a recursion that allows haplotype tests to be applied exactly for small samples; for larger samples, efficient acceptance-rejection algorithms can be applied.

Testing the haplotype frequency distribution for anomalies can be viewed as similar to testing if the “site frequency spectrum,” or the distribution of allele frequencies for segregating sites, fits the standard neutral model (Tajima, 1989). For the site frequency spectrum, an excess of rare variants can reflect population growth or population structure, whereas an excess of common variants can reflect positive selection or balancing selection. Similarly, conditional on the number of segregating sites, an excess of rare haplotypes may reflect population growth or recombination, while an excess of common haplotypes may reflect population structure or positive selection. As is done for tests based on the site frequency spectrum, the HCT, HHT, HNT, and HDT can be used to scan genomes for atypical regions: for the haplotype configuration test, \( c \) is the number of segregating sites, an excess of rare haplotypes may reflect population growth or recombination.

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>( \beta = 0 )</th>
<th>( \beta = 0.1 )</th>
<th>( \beta = 1 )</th>
<th>( \beta = 10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho = 0 )</td>
<td>( C )</td>
<td>1699</td>
<td>1442</td>
<td>469</td>
</tr>
<tr>
<td>( M )</td>
<td>755</td>
<td>635</td>
<td>222</td>
<td>2</td>
</tr>
<tr>
<td>( K )</td>
<td>5481</td>
<td>4702</td>
<td>1764</td>
<td>50</td>
</tr>
<tr>
<td>( H )</td>
<td>781</td>
<td>647</td>
<td>224</td>
<td>2</td>
</tr>
<tr>
<td>( \rho = 0.01 )</td>
<td>( C )</td>
<td>1722</td>
<td>1394</td>
<td>460</td>
</tr>
<tr>
<td>( M )</td>
<td>752</td>
<td>592</td>
<td>230</td>
<td>3</td>
</tr>
<tr>
<td>( K )</td>
<td>5627</td>
<td>4629</td>
<td>1762</td>
<td>46</td>
</tr>
<tr>
<td>( H )</td>
<td>778</td>
<td>612</td>
<td>234</td>
<td>3</td>
</tr>
<tr>
<td>( \rho = 0.1 )</td>
<td>( C )</td>
<td>1776</td>
<td>1381</td>
<td>427</td>
</tr>
<tr>
<td>( M )</td>
<td>766</td>
<td>582</td>
<td>217</td>
<td>2</td>
</tr>
<tr>
<td>( K )</td>
<td>5672</td>
<td>4552</td>
<td>1672</td>
<td>52</td>
</tr>
<tr>
<td>( H )</td>
<td>792</td>
<td>602</td>
<td>222</td>
<td>2</td>
</tr>
<tr>
<td>( \rho = 1 )</td>
<td>( C )</td>
<td>1617</td>
<td>1289</td>
<td>368</td>
</tr>
<tr>
<td>( M )</td>
<td>726</td>
<td>565</td>
<td>182</td>
<td>4</td>
</tr>
<tr>
<td>( K )</td>
<td>5258</td>
<td>4332</td>
<td>1554</td>
<td>36</td>
</tr>
<tr>
<td>( H )</td>
<td>751</td>
<td>581</td>
<td>182</td>
<td>4</td>
</tr>
<tr>
<td>( \rho = 10 )</td>
<td>( C )</td>
<td>674</td>
<td>540</td>
<td>193</td>
</tr>
<tr>
<td>( M )</td>
<td>316</td>
<td>268</td>
<td>97</td>
<td>2</td>
</tr>
<tr>
<td>( K )</td>
<td>2463</td>
<td>2044</td>
<td>795</td>
<td>29</td>
</tr>
<tr>
<td>( H )</td>
<td>320</td>
<td>271</td>
<td>98</td>
<td>2</td>
</tr>
<tr>
<td>( \rho = 100 )</td>
<td>( C )</td>
<td>8</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( M )</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( K )</td>
<td>24</td>
<td>12</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>( H )</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Values are based on \( 10^5 \) accepted genealogies using a point estimate of \( \theta \), estimated from (26) for \( \beta = 0 \) and using \( 2s/\mathbb{E}[L] \) for \( \beta > 0 \), with \( \mathbb{E}[L] \) based on \( 10^5 \) separate simulated genealogies. If the observed configuration is \( c \), the values in the table are \((\times 10^{-3})\): \( \mathbb{P}(C \leq c) \), \( \mathbb{P}[M \geq M(c)] \), \( \mathbb{P}[K \leq K(c)] \), and \( \mathbb{P}[H \leq H(c)] \), where probabilities are estimated from the \( 10^5 \) genealogies. If the observed value of \( C, M, K, \) or \( H \) was located on the boundary of the rejection region, the null hypothesis was not rejected. To reject the null hypothesis at the 0.05 level, numbers shown must be at most 5000 for the test based on \( C \) and 2500 for the other (two-tailed) tests.
that can be placed in \( P \)-values for haplotype tests and other neutrality tests, because in many species, \( \theta \) will no longer need to be estimated from the same data on which the tests are being performed.

ACKNOWLEDGMENTS

We thank A. Hirsh, M. Nordborg, M. Przeworski, and J. Wall for comments. Programs for implementing the simulation-based haplotype configuration test are available from P.M. (pmarjora@usc.edu). This work was supported by a grant from the University of Texas to H.I., a Burroughs-Wellcome Fund Career Award in Biomedical Sciences to N.A.R., NIH grant GM 58897, and Center of Excellence in Genomic work was supported by a grant from the University of Texas to H.I., configurations test are available from P.M. (pmarjora@usc.edu). This

We thank A. Hirsh, M. Nordborg, M. Przeworski, and J. Wall for comments. Programs for implementing the simulation-based haplotype configuration test are available from P.M. (pmarjora@usc.edu). This work was supported by a grant from the University of Texas to H.I., a Burroughs-Wellcome Fund Career Award in Biomedical Sciences to N.A.R., NIH grant GM 58897, and Center of Excellence in Genomic work was supported by a grant from the University of Texas to H.I., configurations test are available from P.M. (pmarjora@usc.edu). This

LITERATURE CITED


Cumulative probability refers to the sum of the probabilities of all configurations that have probability at most $q(c, s, n, \theta)$. The value $\theta = 4.08163$ was chosen using (26). Unlike $\theta = 4.08163$, the values $\theta = 1$ and $\theta = 10$ are unlikely to produce 10 segregating sites.
TABLE 6: Conditional probabilities of haplotype configurations and summary statistics for obtained from simulations.

| Configuration (c) | q(c|s, n, θ) | Cumulative probability | $P[M \geq M(e)]$ | $P[M \leq M(e)]$ | $P[K \leq K(e)]$ | $P[K \geq K(e)]$ | $P[H \leq H(e)]$ | $P[H \geq H(e)]$ |
|------------------|------------|-----------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| (7,0,0,0,0,0,0)  | 0.0437     | 1.0000                | 1.0000          | 0.0437          | 1.0000          | 0.0437          | 1.0000          | 0.0437          |
| (5,1,0,0,0,0,0)  | 0.2297     | 0.0000                | 0.9563          | 0.5306          | 0.9563          | 0.2774          | 0.9563          | 0.2774          |
| (3,2,0,0,0,0,0)  | 0.2198     | 0.7703                | 0.9563          | 0.5306          | 0.7266          | 0.6405          | 0.7266          | 0.4932          |
| (1,3,0,0,0,0,0)  | 0.0374     | 0.1327                | 0.9563          | 0.5306          | 0.3595          | 0.9047          | 0.5068          | 0.6779          |
| (4,0,1,0,0,0,0)  | 0.1473     | 0.3958                | 0.4694          | 0.8629          | 0.7266          | 0.6405          | 0.5068          | 0.6779          |
| (2,1,1,0,0,0,0)  | 0.1547     | 0.5505                | 0.4694          | 0.8629          | 0.3595          | 0.9047          | 0.3221          | 0.8326          |
| (0,2,1,0,0,0,0)  | 0.0144     | 0.0244                | 0.4694          | 0.8629          | 0.0953          | 0.9900          | 0.1674          | 0.8470          |
| (1,0,2,0,0,0,0)  | 0.0159     | 0.0403                | 0.4694          | 0.8629          | 0.0953          | 0.9900          | 0.1530          | 0.9350          |
| (3,0,0,1,0,0,0)  | 0.0721     | 0.2485                | 0.1371          | 0.9696          | 0.3595          | 0.9047          | 0.1530          | 0.9350          |
| (1,1,0,1,0,0,0)  | 0.0319     | 0.0953                | 0.1371          | 0.9696          | 0.0953          | 0.9900          | 0.0650          | 0.9669          |
| (0,0,1,1,0,0,0)  | 0.0027     | 0.0053                | 0.1371          | 0.9696          | 0.0100          | 1.0000          | 0.0331          | 0.9696          |
| (2,0,0,1,0,0,0)  | 0.0231     | 0.0634                | 0.0304          | 0.9953          | 0.0953          | 0.9900          | 0.0304          | 0.9927          |
| (0,1,0,1,0,0,0)  | 0.0026     | 0.0026                | 0.0304          | 0.9953          | 0.0100          | 1.0000          | 0.0073          | 0.9953          |
| (1,0,0,0,1,0,0)  | 0.0047     | 0.0100                | 0.0047          | 1.0000          | 0.0100          | 1.0000          | 0.0047          | 1.0000          |

Bayesian approach with a prior of $\theta \sim \text{unif}[0, 20]$ (Procedure 3)

| Configuration (c) | q(c|s, n, θ) | Cumulative probability | $P[M \geq M(e)]$ | $P[M \leq M(e)]$ | $P[K \leq K(e)]$ | $P[K \geq K(e)]$ | $P[H \leq H(e)]$ | $P[H \geq H(e)]$ |
|------------------|------------|-----------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| (7,0,0,0,0,0,0)  | 0.0373     | 0.1216                | 1.0000          | 0.0373          | 1.0000          | 0.0373          | 1.0000          | 0.0373          |
| (5,1,0,0,0,0,0)  | 0.1951     | 0.7889                | 0.9627          | 0.4852          | 0.9627          | 0.2324          | 0.9627          | 0.2324          |
| (3,2,0,0,0,0,0)  | 0.2112     | 1.0000                | 0.9627          | 0.4852          | 0.7676          | 0.5838          | 0.7676          | 0.4435          |
| (1,3,0,0,0,0,0)  | 0.0417     | 0.1633                | 0.9627          | 0.4852          | 0.4162          | 0.8733          | 0.5565          | 0.6254          |
| (4,0,1,0,0,0,0)  | 0.1402     | 0.4257                | 0.5148          | 0.8292          | 0.7676          | 0.5838          | 0.5565          | 0.6254          |
| (2,1,1,0,0,0,0)  | 0.1680     | 0.5938                | 0.5148          | 0.8292          | 0.4162          | 0.8733          | 0.3746          | 0.7935          |
| (0,2,1,0,0,0,0)  | 0.0159     | 0.0322                | 0.5148          | 0.8292          | 0.1267          | 0.9837          | 0.2065          | 0.8094          |
| (1,0,2,0,0,0,0)  | 0.0198     | 0.0520                | 0.5148          | 0.8292          | 0.1267          | 0.9837          | 0.1906          | 0.9900          |
| (3,0,0,1,0,0,0)  | 0.0798     | 0.2855                | 0.1708          | 0.9557          | 0.4162          | 0.8733          | 0.1906          | 0.9900          |
| (1,1,0,1,0,0,0)  | 0.0424     | 0.2057                | 0.1708          | 0.9557          | 0.1267          | 0.9837          | 0.0910          | 0.9514          |
| (0,0,1,1,0,0,0)  | 0.0043     | 0.0043                | 0.1708          | 0.9557          | 0.0163          | 1.0000          | 0.0486          | 0.9557          |
| (2,0,0,1,0,0,0)  | 0.0323     | 0.0843                | 0.0443          | 0.9933          | 0.1267          | 0.9837          | 0.0443          | 0.9879          |
| (0,1,0,1,0,0,0)  | 0.0053     | 0.0096                | 0.0443          | 0.9933          | 0.0163          | 1.0000          | 0.0121          | 0.9933          |
| (1,0,0,0,1,0,0)  | 0.0067     | 0.0163                | 0.0067          | 1.0000          | 0.0163          | 1.0000          | 0.0067          | 1.0000          |

Fixed-S approach (Hudson et al., 1994)

Cumulative probability refers to the sum of the probabilities of all configurations that have probability at most $q(c|s, n, \theta)$. For the Bayesian approach, empirical probabilities are based on 10,000 accepted genealogies; for the fixed-S approach, empirical probabilities are based on 10,000 genealogies.