Letter to the Editor

Neutrality Tests Based on the Distribution of Haplotypes Under an Infinite-Site Model

Frantz Depaulis and Michel Veuille
Laboratoire d'Ecologie, Université Paris 6, Paris, France

A commonly held opinion in population genetics is that the number of different haplotypes in a random sample of sequences is not an informative measure of polymorphism since, for long DNA sequences, ''all sequences examined may be different from each other'' (Nei 1987, p. 254). Given a sample of n sequences showing S polymorphic sites, Nei considered the case in which $n \ll S$. Let K be the number of haplotypes. Griffiths (1982) studied the joint distribution of K and S as a function of n and θ ($\theta = 4N_e\mu$, where N_e is the diploid effective population size and μ is the neutral mutation rate) in a model without recombination. This analysis considered low values of θ and S compared to n ($S \ll n$), with the result that each new polymorphic site had a high probability of delineating a new haplotype. Thus, K is close to its maximum value ($S + 1$ when S is less than n). In this case, S and K are redundant statistics.

Nei (1987) and Griffiths (1982) considered two opposite situations. In both cases, K was close to its maximum and was not informative. This maximum (hereafter called $K_{\max}$) is limited by either n or $S + 1$, whichever is smallest, and thus depends on the relative magnitudes of two parameters. Between these extreme cases lie realistic conditions of sequence polymorphism data, for which values of n and S are of similar relative magnitudes. In this paper, we investigate the expectation of K (the haplotype number) and H (haplotype diversity) for intermediate values of n and S. In this case, we show that the expectation of K is different from its maximum. This provides useful statistics for comparing observed values with predictions based on neutral mutation-drift equilibrium.

We derive two tests. One, the ''haplotype number test,'' is based on the distribution of K, and the other, the ''haplotype diversity test,'' is based on the distribution of H. They will be referred to below as the K-test and the H-test. The haplotype diversity of the sample takes into account the haplotype frequency distribution. It is defined as:

$$H = 1 - \sum_{i=1}^{K} p_i^2, \qquad (1)$$

where p_i is the frequency of the ith haplotype. These statistics are conditioned on the sample size n and the number of polymorphic sites S under an infinite-sites model. The meaning of the H-test is close to that of Watterson's (1978) homozygosity test, which is based on an infinite-alleles model. However, Watterson's test is conditioned on the number of haplotypes, whereas H is conditioned on S (and n). In a model without recombination, haplotype frequencies conditional on K and θ follow Ewens' (1972) distribution. The K-test is similar to Fu's (1996) W test, except that our test is conditioned on S, while the W test is conditioned on θ. The range of K is between 2 (for $S > 0$) and $K_{\max}$. The range of H is between $2(n - 1)/n^2$ (in the case of two haplotypes of which one is unique) and $1 - 1/n$ (for $K = n$).

For both statistics, significantly low values reveal a structuring of polymorphic sites into a few haplotypes, due to demographic or selective events: population subdivision, recent bottleneck, balanced polymorphism without recombination, incomplete hitchhiking event, or hitchhiking with partial linkage (Hudson and Kaplan 1988; Kaplan, Darden, and Hudson 1988; Takahata 1988; Tajima 1989*b*). High values can result either from an ancient balanced polymorphism with partial linkage, from a starlike genealogy of haplotypes due to population expansion, or from a complete hitchhiking event without recombination (Maynard Smith and Haigh 1974; Hudson and Kaplan 1988; Tajima 1989*b*). Griffiths' (1982) approach was conditional on θ and was not used here. This analysis showed that the expectation of K for a given S substantially depends on θ. We used S, which is directly observed, rather than θ, which needs to be estimated. A coalescent simulation approach was used to derive the expectations and confidence intervals of the two statistics for ranges of parameters usually found in empirical studies (tables 1 and 2). Gene genealogies were performed using standard procedures (Hudson 1990). Time was expressed in units of N_e generations, and S mutations were randomly distributed on the genealogy. Critical values of the discrete statistic K were determined conservatively (e.g., the values indicated in tables 1 and 2 are those of the 250th lowest and highest values obtained out of 10,000 simulations for a two-tailed test with $\alpha = 0.05$). For instance, considering the case $n = 10$ and $S = 10$, at least 2.5% of the runs gave a K value of ≤ 3 and at least 2.5% gave a value of ≥ 8. The 95% confidence interval given in table 1 (3; 8), is therefore conservative.

The expectation of K, given n and S, agreed with the results of Griffiths (1982) for comparable parameter values. The haplotype number K (table 1) is an increasing function of n and S. Haplotype diversity H (table 2) is an increasing function of S, and is also an increasing function of n for most values. However, it behaves as a decreasing function of n for some values, roughly when $n > S$. Tables 1 and 2 indicate the 95% confidence intervals for K and H, respectively. The lower bound of K is above its lowest possible value ($K = 2$) for most

Key words: neutrality test, molecular polymorphism, infinite-site model.

Address for correspondence and reprints: Frantz Depaulis, Université Paris 6, Laboratoire d'Ecologie CNRS-UMR 7625, EPHE, 7 quai Saint Bernard, 75252 Paris cedex 05 France.
E-mail: fdepauli@snv.jussieu.fr.

Mol. Biol. Evol. 15(12):1788–1790. 1998

Table 1
Expectation and Confidence Interval for *K*

	Sample Size (*n*)						
S	5	10	15	20	30	40	50
5	3.40	4.20	4.52	4.71	4.90	5.03	5.11
....	[2; 5]	[2; 6]	[2; 6]	[3; 6]	[3; 6]	[3; 6]	[3; 6]
10	3.92	5.49	6.31	6.79	7.35	7.69	7.97
....	[2; 5]	[3; 8]	[3; 9]	[4; 10]	[4; 10]	[4; 11]	[5; 11]
15	4.17	6.29	7.49	8.24	9.17	9.80	10.21
....	[3; 5]	[4; 9]	[4; 11]	[5; 12]	[5; 13]	[6; 14]	[6; 14]
20	4.33	6.78	8.31	9.35	10.70	11.53	12.15
....	[3; 5]	[4; 9]	[5; 11]	[5; 13]	[6; 15]	[7; 16]	[7; 17]
25	4.44	7.19	9.01	10.26	11.93	13.04	13.86
....	[3; 5]	[4; 10]	[5; 12]	[6; 14]	[7; 17]	[8; 18]	[8; 19]
30	4.52	7.50	9.54	10.99	13.04	14.31	15.33
....	[3; 5]	[5; 10]	[6; 13]	[7; 15]	[8; 18]	[9; 20]	[10; 21]
40	4.62	7.92	10.35	12.16	14.73	16.54	17.89
....	[3; 5]	[5; 10]	[7; 14]	[8; 16]	[9; 20]	[10; 23]	[11; 24]
50	4.69	8.24	10.91	13.02	16.12	18.30	19.91
....	[3; 5]	[6; 10]	[7; 13]	[9; 17]	[11; 21]	[21; 25]	[13; 27]
60	4.73	8.48	11.41	13.73	17.27	19.80	21.74
....	[4; 5]	[6; 10]	[8; 14]	[9; 18]	[12; 23]	[13; 26]	[15; 29]

NOTE.—Numbers are based on 10,000 simulations for a model without intragenic recombination. The upper number is the expectation of *K* (number of different haplotypes). The 95% confidence intervals are indicated in brackets.

Table 2
Expectation and Confidence Interval for *H*

	Sample Size (*n*)						
S	5	10	15	20	30	40	50
5	63	64	62	61	59	57	56
..	[32; 80]	[34; 80]	[34; 77]	[27; 78]	[24; 77]	[23; 77]	[22; 76]
10	70	73	74	73	72	72	71
..	[48; 80]	[48; 86]	[48; 86]	[48; 86]	[44; 86]	[45; 85]	[43; 84]
15	72	78	79	79	78	78	77
..	[56; 80]	[58; 88]	[59; 89]	[57; 89]	[56; 89]	[55; 88]	[54; 88]
20	74	80	82	82	82	82	82
..	[56; 80]	[60; 88]	[61; 90]	[64; 91]	[64; 90]	[62; 90]	[62; 90]
25	75	82	84	84	84	84	84
..	[56; 80]	[66; 90]	[68; 91]	[69; 92]	[68; 92]	[67; 92]	[68; 92]
30	76	83	85	86	86	86	86
..	[56; 80]	[68; 90]	[71; 92]	[72; 92]	[72; 92]	[72; 93]	[72; 93]
40	77	85	87	88	88	89	89
..	[56; 80]	[72; 90]	[75; 92]	[77; 93]	[77; 94]	[77; 94]	[78; 94]
50	77	86	88	89	90	90	90
..	[64; 80]	[76; 90]	[78; 92]	[80; 94]	[81; 95]	[81; 95]	[81; 95]
60	78	86	89	90	91	91	91
..	[72; 80]	[76; 90]	[80; 92]	[82; 94]	[83; 95]	[84; 95]	[84; 95]

NOTE.—All numbers are multiplied by 100. Each position is based on 10,000 simulations in a model without intragenic recombination. The upper number is the expectation of *H* (haplotype diversity of the sample). The 95% confidence intervals are indicated in brackets.

parameter values. More interestingly, the upper bound of *K* is below its maximum possible value ($K_{\max}$) for a wide range of parameters. For instance, this is always observed when $S > 10$, whatever the value of *n.* In other words, even for sequences of substantial length, all sequences do not tend to be different from each other, as intuition would suggest, and each new polymorphic site does not tend to delineate a new haplotype, even in large samples. The same is true for *H,* the boundaries of which are not always equal to its extreme possible values. This is found for all cases in which $n > 10$. This suggests that these statistics can potentially detect haplotype numbers and diversities that are not only lower, but also higher than expected. They can thus be used as two-tailed tests.

In order to assess the usefulness of these tests, we compared them with other tests using available data sets. The two tests were applied to sequence data for the *Su*(*H*) locus in *Drosophila melanogaster* (unpublished data). In this data set, we found seven haplotypes in a sample of 20 sequences encompassing 44 polymorphic sites. The two present statistics were significant ($K = 7$, $P = 0.011$; $H = 0.76$, $P = 0.017$). Other neutrality tests were applied: Tajima's (1989*a*) *D*, Fu and Li's (1993) *D* and *D** test, Kelly's (1997) Z_{nS} test, and the HKA test (Hudson, Kreitman, and Aguadé 1987). None of these tests gave a significant result. They excluded recombination. A recombination model following Hudson (1983) and using the recombination rate estimated from the data ($N_e r = 5.45 \times 10^{-3}$/bp using the method put forward by Hudson 1987) was highly significant ($P < 10^{-4}$) for both the *K*-test and the *H*-test. This estimation of the recombination rate may show considerable error but is close to (and below) the direct estimate based on genetic crosses ($N_e r = 10^{-2}$/bp; Chovnick, Gelbart, and McCarron 1977). A substantial gain of power is thus obtained from models with recombination. A long DNA sequence not only has a large *S*, but is also long from a genetic point of view, thus increasing the contribution of recombination to haplotype formation. A test excluding recombination is conservative for the lower bound of the statistic. If a deviation in this direction is being assessed, a test performed with a conservatively low value of the recombination rate should be used. On the other hand, if a departure toward high values is being assessed, a conservatively high value of the recombination rate should be used.

Another coalescent-based haplotype test has been put forward by Hudson et al. (1994). This test (hereafter HHT: Hudson et al.'s haplotype test) examines whether the frequency of the major haplotypic class is higher than expected. This is a unilateral test, and also, as pointed out by Hudson et al. (1994), an a posteriori test. The three haplotype tests were compared on the *Su*(*H*) data set cited above and on two polymorphism data sets for which the HHT has formerly been used in *D. melanogaster*: the *Sod* locus (Hudson et al. 1994) and the *w* locus (Kirby and Stephan 1995). Which test gave the lowest *P* value depended on the data set and the recombination parameter (results not shown). For instance, in models without recombination, only the *K*-test was nonsignificant for the *Sod* data set, while only the HHT was nonsignificant for the *Su*(*H*) data set. This suggests that the three haplotype tests, although similar, are not redundant. As discussed earlier, the usefulness of these tests depends on the relative values of *n, S,* and the recombination rate, but the tests are not limited to narrow ranges of parameter values. In the case of substantial recombination rates or for long sequences, a sliding-window approach can be used, as for the HHT (Kirby and Stephan 1995). Concerning models without recombination, a theoretical treatment by Fu (1996, 1997) examined the power of two tests, *W* and F_S, which are close to K, based on the number of haplotypes and on Ewens' (1972) distribution. He found that such tests would generally do better than other available neutrality tests. For example, Fu's W test is significant ($P = 0.01$) when applied to $Su(H)$ data. Their main drawback is their dependence on an estimate of θ, for which several estimators, each with different statistical properties, are available. This approach leads to an ''achieved level of significance,'' which may be quite different from the reference error risk (Fu 1996). This tends to support our choice of primary data, n and S, for the K- and H-tests. The question of the power of the present tests against the various alternative hypotheses remains to be addressed more specifically.

In summary, this study shows that the haplotype number (K) and the haplotype diversity (H) of DNA polymorphism data are not trivially determined by the sequence length or the sample size. They are simple and informative statistics for describing the distribution of haplotypes under an infinite-sites model. Confidence intervals conditional on the sample size (n) and the number of polymorphic sites (S) are included in tables 1 and 2. These are ideally designed for data sets from nonrecombining sequences. For data sets with recombination, these values are conservative for the lower bound, but more exact critical-values distribution can be easily obtained from coalescent simulations with recombination. The two neutrality tests based on these distributions, the K-test and the H-test, appear in several instances to be superior to other available tests in detecting departure from neutrality in DNA polymorphism data.

Acknowledgments

This work was financially supported by CNRS. We thank Frédéric Austerlitz, Matthew Cobb, Martha Hamblin, Bernard Jung-Muller, Etienne Klein, and anonymous reviewers for helpful comments on the manuscript.

LITERATURE CITED

Chovnick, A., W. Gelbart, and M. McCarron. 1977. Organization of the *Rosy* locus in *Drosophila melanogaster*. Cell **11**:1–10.

Ewens, W. J. 1972. The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3**:87–112.

Fu, Y. X. 1996. New statistical tests of neutrality for DNA samples from a population. Genetics **143**:557–570.

———. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147**:915–925.

Fu, Y. X., and W. H. Li. 1993. Statistical tests of neutrality of mutations. Genetics **133**:693–709.

Griffiths, R. C. 1982. The number of alleles and segregating sites in a sample from the infinite-alleles model. Adv. Appl. Prob. **14**:225–239.

Hudson, R. R. 1983. Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23**:183–201.

———. 1987. Estimating the recombination parameter in a finite population model without selection. Genet. Res. Camb. **50**:245–250.

———. 1990. Gene genealogies and the coalescent process. Pp. 1–44 in D. Futuyma and J. Antonovics. Oxford surveys in evolutionary biology. Oxford University Press, Oxford.

Hudson, R. R., K. Bailey, D. Skarecky, J. Kwiatowski, and F. J. Ayala. 1994. Evidence for a positive selection in the Superoxide Dismutase (*Sod*) region of *Drosophila melanogaster*. Genetics **136**:1329–1340.

Hudson, R. R., and N. L. Kaplan. 1988. The coalescent process in models with selection and recombination. Genetics **120**:831–840.

Hudson, R. R., M. Kreitman, and M. Aguadé. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics **116**:153–159.

Kaplan, N. L., T. Darden, and R. R. Hudson. 1988. The coalescent process in models with selection. Genetics **120**: 819–829.

Kelly, J. K. 1997. A test of neutrality based on interlocus associations. Genetics **146**:1197–1206.

Kirby, D. A., and W. Stephan. 1995. Haplotype test reveals departure from neutrality in a segment of the *white* gene of *Drosophila melanogaster*. Genetics **141**:1483–1490.

Maynard Smith, J., and J. Haigh. 1974. The hitch-hiking effect of a favorable gene. Genet. Res. Camb. **23**:23–35.

Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

Tajima, F. 1989*a*. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**: 585–595.

———. 1989*b*. The effect of change in population size on DNA polymorphism. Genetics **123**:597–601.

Takahata, N. 1988. The n coalescent in two partially isolated diffusion populations. Genet. Res. Camb. **52**:213–222.

Watterson, G. A. 1978. The homozygosity test of neutrality. Genetics **88**:405–417.

Pierre Capy, reviewing editor

Accepted September 2, 1998