Letter to the Editor

On a Test of Depaulis and Veuille

Lada Markovtsova, Paul Marjoram,† and Simon Tavaré*†‡*

*Department of Mathematics, †Biostatistics Division, Department of Preventive Medicine, and ‡Program in Molecular Biology, Department of Biological Sciences, University of Southern California

In a recent letter to this journal, Depaulis and Veuille (1998) discussed two possible tests of neutrality, the ''haplotype number test'' and the ''haplotype diversity test.'' They present in their tables 1 and 2 means and percentage points of the distribution of the number K_n of haplotypes and the sample heterozygosity $H_n = 1 - \sum_{i=1}^{K_n} p_i^2$, where $p_1, \ldots, p_{K_n}$ are the relative frequencies of those haplotypes in a sample of size n for different values of the number of segregating sites s observed in the data. They assume a neutral infinitely-many-sites model of mutation with no recombination. These percentage points were found by repeatedly simulating a random coalescent tree with n tips, randomly distributing s mutations on the tree, and calculating the observed values of K_n and H_n. See Hudson (1990) for a description of how such simulations can be performed.

Depaulis and Veuille's (1998) procedure produces observations whose distribution is independent of the underlying neutral mutation rate θ. The authors present the method as though the resulting simulated values had the conditional distribution of K_n and H_n given $S_n = s$, respectively. However, this is not true, as the following argument shows. Denote the coalescent tree by Λ and the coalescence times by $T = (T_n, \ldots, T_2)$, so that T_j is the time for which there are j distinct ancestors in the tree Λ. We see that for mutation rate θ, the conditional distribution of (K_n, H_n) given $S_n = s$ can be represented as

$$\mathbb{P}_\theta(K_n = k, H_n = h \,|\, S_n = s) = \sum_\lambda \int_t \mathbb{P}_\theta(K_n = k, H_n = h \,|\, \Lambda = \lambda, T = t, S_n = s) \times f_\theta(\lambda, t \,|\, s)\, dt, \quad (1)$$

where $f_\theta(\lambda, t \,|\, s)$ is the joint conditional density of (Λ, T) given $S_n = s$. Because of the Poisson nature of the mutation process, the first term under the integral signs in equation (1) does not depend on θ. It follows that in order to simulate observations from the joint conditional distribution of (K_n, H_n) given $S_n = s$, one first simulates from the conditional distribution of (Λ, T) given $S_n = s$ and then randomly distributes the s mutations over the resulting tree and calculates the values of K_n and H_n. Notice that Depaulis and Veuille's (1998) procedure simulates from the unconditional distribution of (Λ, T) instead of the conditional distribution of (Λ, T) given $S_n = s$. The joint distribution of (K_n, S_n) in the case of constant population size is discussed by Griffths (1982), and that in the variable population size case is discussed by Griffths and Tavaré (1996).

Depaulis and Veuille (1998) suggested that the percentage points of their statistics could be used as a test of neutrality: for a given sample size n and observed value of s, one compares the observed values of K_n and H_n with the given 95% credible intervals. Values falling outside those intervals lead to rejection of neutrality. Given that the true conditional distributions of K_n and H_n in fact depend on the unknown mutation rate θ, it is likely that the power of this test varies dramatically as a function of θ for given n and s. To assess this hypothesis, we simulated observations from the true joint conditional distribution using equation (1) and then estimated the probability that either statistic would fall outside the limits given by Depaulis and Veuille. This gave an empirical estimate of the probability that their test would reject neutrality for different values of θ.

We used a Markov chain Monte Carlo (MCMC) approach to simulate observations from the conditional distribution of (Λ, T) given $S_n = s$. MCMC methods produce correlated samples, but these samples may be made approximately independent by sampling the output at widely spaced intervals. The results below were generated using the approach in Markovtsova, Marjoram, and Tavaré (2000). An alternative approach is the rejection algorithm of Tavaré et al. (1997). Table 1 shows the fraction of 10,000 observations that fell inside the DV nominal 95% intervals for three different scenarios. As noted by Fu and Li (1993), S_n is not a sufficient statistic for θ, so the fraction of observations that fall within the DV nominal 95% intervals varies greatly. It appears that if the true θ value is well supported by the data, the test of neutrality based on the DV intervals will work well. However, if the true θ value is not well supported by the data, the test will be inaccurate, leading to incorrect rejections of neutrality. For further discussion, see Wall and Hudson (2001).

As we have seen, the power of the DV test depends on the unknown parameter θ. Even if the mutation rate is known, the compound parameter θ still depends on the underlying effective population size at the time of sampling, and this is not known in practice. Depaulis and Veuille (1998) discuss a data set from the *Su(H)* locus in *Drosophila melanogaster* for which $n = 20$, $s = 44$. The observed values of K_{20} and H_{20} were 7 and 0.76, respectively. The nominal P values were estimated to be 0.011 and 0.017, respectively. We used the simulation approach outlined above to find empirical estimates of these P values for different values of θ, using 10,000 simulated values once more. We used values of

Key words: Markov chain Monte Carlo, coalescent.

Address for correspondence and reprints: Simon Tavaré, Program in Molecular Biology, Department of Biological Sciences, SHS 172, University of Southern California, Los Angeles, California 90089-1340. E-mail: stavare@gnome.usc.edu.

Mol. Biol. Evol. 18(6):1132–1133. 2001

Table 1
Fraction (%) of 10,000 Observations Falling Within the DV Interval

	95% DV Interval	θ = 1	θ = 10	θ = 50	θ = 100
$n = 10, s = 10$					
K_n	[3, 8]	96.1	98.3	96.2	95.7
H_n	[0.48, 0.86]	91.4	97.8	96.1	95.7
$n = 20, s = 40$					
K_n	[8, 16]	17.0	98.0	83.3	75.2
H_n	[0.77, 0.93]	30.6	97.1	87.2	81.0
$n = 50, s = 50$					
K_n	[13, 27]	1.9	98.6	56.9	34.1
H_n	[0.81, 0.95]	17.5	97.3	79.0	66.8

θ = 1, 5, 12.4, 50.0, and 100 for illustration; the value 12.4 corresponds to Watterson's (1975) segregating-sites estimator. Results are shown in table 2.

From table 2, we see that for θ in the range 12.4 or larger, the data are highly unlikely under a neutral scenario. However, for a range of smaller θ values, including, for example, θ = 5, the data become much more likely. Thus, the ability to reject the supposed neutral scenario depends on the true value of θ. It should be noted that one cannot reject neutrality on the basis of this test; rather, one can reject the model upon which the test is based. As Depaulis and Veuille (1998) note, this model does not include recombination (although it is easy to alter it to do so); neither does it include any population demographics such as stratification. If the model is rejected, any of these missing factors could be the cause, not necessarily the assumption of neutrality.

Computer programs that implement both the MCMC approach and the rejection method to generate observations from the joint conditional distribution of, for example, (K_n, H_n) given the value of S_n for a given distribution for θ under a variety of demographic scenarios can be obtained from the authors.

Table 2
***P* Values for *Drosophila melanogaster* Data**

	θ = 1	θ = 5	θ = 12.4	θ = 50	θ = 100
K.	0.346	0.135	0.005	0.000	0.000
H.	0.345	0.121	0.013	0.000	0.000

Acknowledgments

We thank Y.-X. Fu and an anonymous reviewer for helpful comments. We were supported in part by NSF grant DBI95-04393 and NIH grant GM 58897.

LITERATURE CITED

DEPAULIS, F., and M. VEUILLE. 1998. Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol. Biol. Evol. **15**:1788–1790.

FU, Y. X., and W. H. LI. 1993. Maximum likelihood estimation of population parameters. Genetics **134**:1261–1270.

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.

GRIFFITHS, R. C., and S. TAVARÉ. 1996. Monte Carlo inference methods in population genetics. Math. Comput. Modelling **23**:141–158.

HUDSON, R. R. 1990. Gene genealogies and the coalescent process. Pp. 1–44 *in* D. FUTUYMA and J. ANTONOVICS, eds. Oxford surveys in evolutionary biology. Vol. 7. Oxford University Press, Oxford, England.

MARKOVTSOVA, L., P. MARJORAM, and S. TAVARÉ. 2000. The effects of rate variation on ancestral inference in the coalescent. Genetics **156**:1427–1436.

TAVARÉ, S., D. BALDING, R. C. GRIFFITHS, and P. DONNELLY. 1997. Inferring coalescence times for molecular sequence data. Genetics **145**:505–518.

WALL, J. D., and R. R. HUDSON. 2001. Coalescent simulations and statistical tests of neutrality. Mol. Biol. Evol. **18**:1134–1135.

WATTERSON, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7**:256–276.

YUN-XIN FU, reviewing editor

Accepted January 29, 2001