A New Statistic for Detecting Genetic Differentiation

Richard R. Hudson

Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637

Manuscript received February 15, 2000
Accepted for publication May 1, 2000

ABSTRACT

A new statistic for detecting genetic differentiation of subpopulations is described. The statistic can be calculated when genetic data are collected on individuals sampled from two or more localities. It is assumed that haplotypic data are obtained, either in the form of DNA sequences or data on many tightly linked markers. Using a symmetric island model, and assuming an infinite-sites model of mutation, it is found that the new statistic is as powerful or more powerful than previously proposed statistics for a wide range of parameter values.

DETECTING genetic differentiation of subpopulations is an important problem in several areas of population biology, including areas of evolutionary genetics, ecology, and conservation biology. When data are obtained from two or more localities in the form of allele frequencies at one or more unlinked loci, standard chi-square tests (or likelihood-ratio tests) of homogeneity are appropriate (Workman and Niswander 1970) and can be quite powerful for detecting differentiation. Even when the expected counts in some cells are small, permutation methods can be utilized to give good results (Lewontin and Felsenstein 1965; Roff and Bentzen 1989). If the data consist of DNA sequences, or haplotyping at two or more linked sites, the same methods can be employed, if distinct sequences or haplotypes are treated as alleles. However, if the haplotype diversity is very high and the sample sizes are small, most haplotypes may appear in the sample only once and the methods based on haplotype frequencies will have low power and, in extreme cases, can become completely useless. Using these methods, longer sequences, which must contain more information, can result in lower power than short sequences. This problem is most severe with small samples and long sequences. To handle these kinds of data, Hudson *et al.* (1992) proposed the use of sequence-based statistics in the permutation tests. These sequence-based statistics utilize information on the numbers of differences between haplotypes and not just the frequencies of the haplotypes. The particular sequence-based statistics considered by Hudson *et al.* (1992) were shown to be more powerful than the chi-square statistic when haplotype diversity was very high, but were found to be relatively weak when the diversity was low. Thus, for low diversity samples, the chi-square statistic (or a likelihood-ratio statistic) would appear to be best, but, for high diversity samples, the sequence-based statistics should be used. Unfortunately, there are no absolute criteria known for when the chi-square statistic should be employed and when the sequence-based statistics should be used. It would be desirable to have a single statistic that performs well at all levels of diversity. In this note, a new sequence-based statistic is introduced that appears to have this property. Under a symmetric two-island model with mutations occurring according to the infinite-sites model, this new statistic is found to be as powerful or more powerful than other statistics that have been proposed for detecting genetic differentiation. This superior power is found over a wide range of haplotype diversity.

The new statistic, referred to as the nearest-neighbor statistic (S_{nn}), is a measure of how often the "nearest neighbors" (in sequence space) of sequences are from the same locality in geographic space. This is made more precise below. The statistic is applicable when genetic data are collected on individuals sampled from two or more localities. It is assumed that haplotypic data are obtained, either in the form of DNA sequences or data on many tightly linked markers.

To define S_{nn}, it is helpful to first establish some notation. For concreteness, suppose the data collected are mitochondrial sequences obtained from n individuals, some of which are from locality 1 and some from locality 2. (The statistic automatically generalizes to more localities.) We assume all sequences are the same length with no gaps. Arbitrarily number the individuals from 1 to n, and denote the sequence of individual i by s_i. Let d_{ij} equal the number of nucleotide sites at which s_i differs from s_j. Focus on a particular individual, say, individual k, and let m_k denote the minimum of $\{d_{kj}\}$, $j = 1, 2, \ldots, k - 1, k + 1, \ldots, n$. Thus, m_k is the distance to the nearest neighbor(s) of individual k. (Neighbor here reflects closeness in sequence space, not in geographic space.) Let T_k equal the number of individuals for which $d_{kj} = m_k$, again for fixed k and $j \neq k$. T_k is the number

Corresponding author: Richard R. Hudson, Department of Ecology and Evolution, University of Chicago, 1101 E. 57th St., Chicago, IL 60637. E-mail: rr-hudson@uchicago.edu

TABLE 1

Power of tests (cases examined by Hudson *et al.* 1992)

					Power				
n_1	n_2	$4Nu$	$4Nc$	$4Nm$	K_S*	Z*	H_S	χ^2	S_{nn}
35	5	5.0	0.0	2.0	0.58	0.62	0.62	0.78	0.77
30	10				0.79	0.83	0.83	0.91	0.94
25	15				0.87	0.90	0.90	0.96	0.98
20	20				0.88	0.92	0.91	0.97	0.98
10	10			5.0	0.32	0.34	0.32	0.36	0.42
			20.0		0.46	0.44	0.21	0.21	0.46
15	15		0.0		0.47	0.52	0.52	0.63	0.66
			20.0		0.70	0.66	0.54	0.60	0.74
25	25		0.0	1.0	0.99	1.00	0.99	1.00	1.00
			20.0		1.00	1.00	0.99	1.00	1.00
			0.0	2.0	0.94	0.96	0.95	0.99	1.00
			20.0		1.00	0.99	0.98	1.00	1.00
			0.0	5.0	0.69	0.75	0.79	0.91	0.92
			20.0		0.90	0.88	0.87	0.95	0.96
			0.0	10.0	0.41	0.46	0.53	0.68	0.69
			20.0		0.68	0.63	0.70	0.81	0.81
50	50		0.0	5.0	0.91	0.95	0.97	1.00	1.00
			20.0		1.00	0.99	1.00	1.00	1.00
			0.0	10.0	0.71	0.78	0.85	0.96	0.97
			20.0		0.95	0.93	0.97	1.00	1.00
25	25	0.156	0.0	5.0	0.17	0.17	0.16	0.19	0.22
			0.624		0.18	0.18	0.17	0.19	0.23
		0.313	0.0		0.30	0.29	0.28	0.33	0.37
			1.25		0.31	0.30	0.30	0.35	0.38
		0.625	0.0		0.41	0.41	0.40	0.50	0.53
			2.5		0.46	0.46	0.44	0.56	0.57
		1.25	0.0		0.53	0.54	0.53	0.67	0.69
			5.0		0.62	0.64	0.64	0.78	0.79
		2.5	0.0		0.61	0.66	0.68	0.83	0.84
			10.0		0.78	0.77	0.79	0.91	0.91
		5.0	0.0		0.69	0.75	0.79	0.91	0.92
			20.0		0.90	0.88	0.87	0.95	0.96
		10.0	0.0		0.76	0.83	0.87	0.95	0.96
			40.0		0.98	0.96	0.85	0.90	0.98
		15.0	0.0		0.78	0.86	0.88	0.96	0.97
			60.0		0.99	0.98	0.74	0.76	0.99

For each row of this table, 4000 independent samples were generated under a symmetric two-island model. For each of these samples, 4000 random permutations were carried out to estimate the P value of each of the statistics for the sample. n_1 and n_2 are the sample sizes from locality one and locality two, respectively. N is the population size of each subpopulation. u is the neutral mutation rate per generation. c is the per generation recombination rate between the ends of the segment sequenced. m is the migration fraction per generation. K_S*, Z*, and H_S are the sequence-based statistics considered by Hudson *et al.* (1992). χ^2 is the chi-square statistic and S_{nn} is the nearest-neighbor statistic. The power estimates are the proportion of samples with estimated P value <0.05.

of nearest neighbors of individual k. And let W_k equal the number of individuals with $d_{kj} = m_k$, that are from the same locality as individual k. In other words, W_k is the number of nearest neighbors to individual k that are from the same locality as individual k. Now define $X_k = W_k / T_k$. Thus, X_k is the fraction of nearest neighbors of individual k that are from the same locality as individual k. Thus, if individual k has only a single nearest neighbor, then X_k is one if the nearest neighbor is from the same locality as individual k, and X_k is zero if the nearest neighbor is from a different locality. The statistic S_{nn} is simply the average of the X_k:

$$S_{nn} = \sum_{j=1}^{n} X_j / n.$$

S_{nn} is a measure of how often the nearest neighbors of sequences are found in the same locality. If a population is strongly structured, one expects to find the nearest neighbor of a sequence in the same locality. Thus, S_{nn} is expected to be near one when the populations at the

TABLE 2

Power of tests in very small sample sizes

							Power				
n_1	n_2	$4Nu$	$4Nc$	$4Nm$	SS	Het	K_S*	Z*	H_S	χ^2	S_{nn}
6	6	0.75	0.0	2.0	5.0	0.62	0.20	0.20	0.18	0.18	0.28
		1.0			6.8	0.69	0.23	0.24	0.21	0.21	0.30
		2.0			13.6	0.82	0.33	0.33	0.26	0.27	0.36
		4.0			27.0	0.90	0.39	0.40	0.24	0.24	0.40
		10.0			67.4	0.96	0.43	0.43	0.09	0.09	0.43
		0.75	0.75		5.1	0.65	0.19	0.19	0.18	0.17	0.26
		1.00	1.0		6.7	0.72	0.25	0.25	0.22	0.22	0.30
		2.0	2.0		13.7	0.85	0.36	0.36	0.27	0.27	0.38
		4.0	4.0		27.2	0.93	0.43	0.42	0.18	0.18	0.41
		10.0	10.0		67.6	0.97	0.53	0.51	0.03	0.03	0.44
10	10	0.75	0.0	5.0	5.5	0.61	0.18	0.18	0.17	0.20	0.24
		1.0			7.5	0.68	0.22	0.23	0.21	0.25	0.27
		2.0			15.0	0.81	0.26	0.27	0.24	0.30	0.34
		4.0			29.7	0.90	0.31	0.34	0.30	0.35	0.40
		10.0			74.5	0.95	0.37	0.40	0.30	0.32	0.46
		0.75	0.75		5.6	0.63	0.19	0.19	0.17	0.21	0.24
		1.0	1.0		7.47	0.71	0.23	0.23	0.22	0.26	0.28
		2.0	2.0		14.8	0.84	0.30	0.31	0.29	0.33	0.37
		4.0	4.0		29.5	0.92	0.35	0.35	0.30	0.33	0.41
		10.0	10.0		74.1	0.97	0.46	0.42	0.20	0.20	0.45

Simulations carried out as for Table 1. *SS* is the mean number of polymorphic sites in the samples. Het is the average haplotype diversity of the samples. Other quantities are as defined in Table 1.

two localities are highly differentiated and near one-half when the populations at the two localities are part of the same panmictic population (and sample sizes from the two localities are equal). To assess whether S_{nn} is significantly large for a particular sample (indicating that the populations at the two localities are differentiated), the usual permutation scheme is applied to estimate a P value. Specifically, a permutation consists of randomly reassigning sequences to localities, so that the number of sequences from each locality is always the same as in the original sample. The proportion of permuted samples with S_{nn} larger than or equal to the observed value is the estimated P value.

To assess the power of permutation tests using S_{nn} to detect geographic differentiation, the same symmetric two-island model considered by Hudson *et al.* (1992) was used. The parameters of this model are N, the island population size, u, the neutral mutation rate, c, the recombination rate between the ends of the segment sequenced, and m, the migration fraction per generation. An infinite-sites model was assumed (and thus no recurrent mutations occur in these simulations.) The results of these simulations are shown in Tables 1 and 2. Table 1 shows the results for all parameter values and sample sizes considered by Hudson *et al.* (1992). In Table 2, more results for small sample sizes are given. For comparison, the power of the permutation tests based on the chi-square statistic (χ^2) and on K_S*, Z*, and H_S are also shown in the tables. The statistics K_S*, Z*, and H_S were the most powerful sequence-based statistics found by Hudson *et al.* (1992).

In Table 1, we find that S_{nn} has equal or higher power than the χ^2 statistic in all cases except one. (The exception is the first case in Table 1 in which the power of S_{nn} was 0.77 while the power of χ^2 was 0.78, a very small difference.) For most cases in this table, S_{nn} has equal or only slightly higher power than the test based on χ^2. However, in cases with small sample sizes ($n_1 = n_2 = 10$ or 15), especially with recombination, there is substantially higher power with the S_{nn} statistic. (In the case with $n_1 = n_2 = 10$ and $4Nc = 20$, the power with S_{nn} is 0.46, while the power with χ^2 is 0.21.) These results motivated us to look at more cases with small sample size, which are shown in Table 2.

For samples of size 6 from each locality, the S_{nn} statistic is substantially more powerful than the χ^2 statistic at all levels of variation examined (see Table 2). For samples of size 10 from each locality, S_{nn} is only slightly more powerful than χ^2 at low levels of variation, but at higher levels of variation, S_{nn} has very much higher power than χ^2. In contrast to the chi-square statistic, higher mutation rates (longer sequences) always lead to more power using the nearest-neighbor statistic, which accords with the intuition that longer sequences should provide more information. With low to moderate levels of variation, the S_{nn} statistic is more powerful than the sequence-based statistics of Hudson *et al.* However, with the small sample sizes considered in Table 2, it appears that K_S*

and Z^* may have slightly higher power than S_{nn} when levels of variation are very high. (See the case $n_1 = n_2 = 6$ and $4Nu = 4Nc = 10$.)

Summarizing, we find that among the statistics tested, S_{nn} is the most powerful statistic, or nearly as powerful as the best statistic, under all conditions examined. It should be emphasized, however, that all assessments of power were carried out with a symmetric two-island model and assuming that mutations occur according to an infinite-sites model (in which no multiple hits occur). Other models may lead to different conclusions. The use of S_{nn} eliminates the need to establish criteria for when to use an "allele" frequency-based statistic and when to use a sequence-based statistic. This statistic may also be of use in testing for genetic differences between samples in case-control studies, though these usually consist of diploid data in large samples.

Source code (in the language C) for a program that carries out the test on Unix or Linux machines is in a file, snn.c, available at http://home.uchicago.edu/~rhudson1.

LITERATURE CITED

Hudson, R. R., D. D. Boos and N. L. Kaplan, 1992 A statistical test for detecting geographic subdivision. Mol. Biol. Evol. **9:** 138–151.

Lewontin, R. C., and J. Felsenstein, 1965 The robustness of homogeneity tests in 2 × N tables. Biometrics **21:** 19–33.

Roff, D. A., and P. Bentzen, 1989 The statistical analysis of mitochondrial DNA polymorphisms: chi 2 and the problem of small samples. Mol. Biol. Evol. **6:** 539–545.

Workman, P. L., and J. D. Niswander, 1970 Population studies on southwestern Indian tribes. II. Local genetic differentiation in the Papago. Am. J. Hum. Genet. **22:** 24–49.

Communicating editor: M. Slatkin