Theoretical Population Biology **60**, 107–116 (2001)
doi:10.1006/tpbi.2001.1522, available online at http://www.idealibrary.com on IDEAL®

Exponential Distance Statistics to Detect the Effects of Population Subdivision

Nancy M. Sundell* and Richard T. Durrett†

*Applied Mathematics and †Mathematics, Cornell University, Ithaca, New York 14853

Received September 20, 2000

Statistical tests are needed to determine whether spatial structure has had a significant effect on the genetic differentiation of subpopulations. Here we introduce a new family of statistics based on a sum of an exponential function of the distances between individuals, which can be used with any genetic distance (e.g., nucleotide differences, number of nonshared alleles, or separation on a phylogenetic tree). The power of the tests to detect genetic differentiation in Wright–Fisher island models and stepping stone models was calculated for various sample sizes, rates of migration and mutation, and definitions of spatial neighborhoods. We found that our new test was in some cases more powerful than the K_s^* statistic of Hudson *et al.* (*Mol. Biol. Evol.* 9, 138–151, 1992), but in all cases was slightly less powerful than both a traditional χ^2 test without lumping of rare haplotypes and the S_{nn} test of Hudson (*Genetics* 155, 2011–2014, 2000). However, when we applied our new tests to three data sets, we found in some cases highly significant results that were missed by the other tests.

INTRODUCTION

In many situations one wants to determine whether the geographical structure of a species contributes to the observed genetic variation. Many types of data can be used to study this question including restriction site mappings, nucleotide sequences (mitochondrial and nuclear), and microsatellites. One of the original methods used to measure the extent of genetic differentiation of subpopulations is Wright's (1951) F_{ST}, the calculation of which involves the mean and variance of gene frequencies. Nei later defined the statistic G_{ST} to be the ratio of intersubpopulational gene diversity to the total gene diversity (Nei, 1973). In general, this is an extension of F_{ST} to the case of multiple alleles (see Takahata and Nei, 1984). Both F_{ST} and G_{ST} are called haplotype statistics, they only use information about gene or haplotype frequencies. One drawback of these statistics is that they fail to consider the number of differences between pairs of haplotypes.

With the availability of nucleotide level data came the introduction of the statistics γ_{ST} (Nei, 1982), N_{ST} (Lynch and Crease, 1990), and, for microsatellites, Slatkin's (1995) R_{ST}. Hudson (2000) recently introduced the statistic S_{nn} which is a measure of how often the nearest neighbors of a sequence are found at the same location. Other tests include the traditional χ^2 test based on allele frequencies (Nei, 1987, p. 227) and the exact probability tests which are based on the classical Fisher test for $R \times C$ contingency tables (Raymond and Rousset, 1995; Goudet *et al.*, 1996). Hudson *et al.* (1992) introduced a permutation approach for determining significance levels for these tests, thereby increasing their power.

Another related technique is the analysis of molecular variance, which examines correlations of haplotypic diversity among demes (Excoffier *et al.*, 1992; Michalakis and Excoffier, 1996). The method involves constructing a hierarchical analysis of molecular variance directly from the matrix of squared distances between all pairs of haplotypes. The main advantage of this method is that the distance matrix can be constructed using assumptions about the evolutionary process causing the haplotype differences.

0040-5809/01 $35.00

One disadvantage of the methods mentioned above is that they all require an *a priori* definition of hierarchical structure. Holsinger and Mason-Gamer (1996) define a new statistic, similar to Nei's G_{ST}, which can be used to group populations based on estimates of the average time to coalescence for pairs of haplotypes. They then construct a tree depicting the pattern of genetic differentiation among subpopulations and test the statistical significance of the groupings.

Other methods utilizing phylogenies of individuals or alleles have been developed to determine the amount of gene flow between populations. Slatkin and Maddison (1989) outline a method using the phylogeny of nonrecombining segments of DNA. After constructing the phylogeny, they estimate the minimum number of migration events, the expected value of which is shown to be a function of Nm. While this method was developed to study the amount of gene flow, it can be used to infer the presence of genetic differentiation as well. One difficulty with their approach is that in order to use it, the underlying population structure must be well approximated by an island model.

Templeton (1998) developed a method incorporating evolutionary genealogical information into the calculation of a statistic. Using a haplotype tree, he defines a nested series of branches (clades) which are then used to analyze the spatial distribution of genetic variation. This analysis involves the comparison of the average distance of individuals in a particular nested clade from the geographical center of that nested clade to the average distance of individuals to the geographical center of the entire clade (Turner *et al.*, 2000).

STATISTICS

Here we introduce new statistics which can be used on any data for which we can define and compute a measure of distance between individuals. This includes phylogenetic trees, restriction site mappings, nucleotide sequences, and microsatellites.

We begin by noting that many statistics used for detecting the influence of population structure have the general form

$$\sum_{k=1}^{L} w_k \sum_{i=1}^{n_k-1} \sum_{j=i+1}^{n_k} F(d(x_{ik}, x_{jk})), \tag{1}$$

where L is the number of subpopulations, n_k is the number of individuals in subpopulation k, w_k is a weighting function defined for each subpopulation, $F(x)$ is a function, and $d(x_{ik}, x_{jk})$ is the distance from individual i in subpopulation k to individual j in subpopulation k. The definition of the distance function, $d(x_{ik}, x_{jk})$, depends on the type of data being used. For a phylogenetic tree this distance could be defined as the number of internal nodes on the shortest path connecting two individuals or the sum of the lengths of the branches along this same path. For a set of nucleotide sequences the distance could be defined as the number of pairwise differences between the sequences.

We propose a family of statistics with $F(x) = u^{-x}$ where $1 < u < \infty$. For simplicity we take $w_k = 1$ for all subpopulations k, but as in Hudson *et al.* (1992), the performance could be improved by optimizing the weights. Our statistics have the form

$$D_u = \sum_{k=1}^{L} \sum_{i=1}^{n_k-1} \sum_{j=i+1}^{n_k} u^{-d(x_{ik}, x_{jk})}. \tag{2}$$

If $u = 1$, we see that D_u is equal to the total number of pairs of individuals in the same subpopulation. Because this value gives us no information about the genetic population structure we require that $u > 1$. It is, however, interesting to look at the limit as u approaches 1. If we let $u = 1+\varepsilon$ with ε small, then $(1+\varepsilon)^{-d(x_{ik}, x_{jk})} \approx 1-\varepsilon d(x_{ik}, x_{jk})$, and the statistic becomes

$$D_{1+\varepsilon} \approx \sum_{k=1}^{L} \sum_{i=1}^{n_k-1} \sum_{j=i+1}^{n_k} (1-\varepsilon d(x_{ik}, x_{jk})). \tag{3}$$

Since $\sum_{k=1}^{L} \sum_{i=1}^{n_k-1} \sum_{j=i+1}^{n_k} 1$ is a constant, this suggests that we define a new statistic that drops this constant term and eliminates the $-\varepsilon$ multiplier from the second term:

$$D_1 = \sum_{k=1}^{L} \sum_{i=1}^{n_k-1} \sum_{j=i+1}^{n_k} d(x_{ik}, x_{jk}). \tag{4}$$

This can be written in the general form (1) by defining $w_k = 1$ and $F(x) = x$.

This new D_1 statistic is similar to the weighted K_s statistic discussed by Hudson *et al.* (1992). Letting $N = \sum n_k$ be the total number of individuals sampled we can define their statistic

$$K_s = \sum_{k=1}^{L} \frac{1}{(n_k-1)} \sum_{i=1}^{n_k-1} \sum_{j=i+1}^{n_k} \frac{2d(x_{ik}, x_{jk})}{N}. \tag{5}$$

This can be written in the general form (1) by defining $w_k = 1/(n_k-1)$ and $F(x) = 2x/N$. We see that the K_s statistic is the D_1 statistic multiplied by a factor involving the sample sizes of the subpopulations. Hudson

et al. (1992) also define the more powerful statistic, K_s^*, which substitutes $\log(1+d(x_{ik}, x_{jk}))$ for $d(x_{ik}, x_{jk})$. This corresponds to $F(x) = (2/N)\cdot\log(1+x)$. They found that the K_s^* statistic was more powerful when using the weight $w_k = \frac{n_k - 2}{n_k(n_k-1)}$. (This formula is slightly different from that of Hudson *et al.* (1992) and has been rewritten to accommodate the general form (1) for sequence statistics used here. Neglecting the constant, the statistic being calculated is the same.)

It is also interesting to look at the statistic D_u in the limit as u approaches positive infinity. As $u \to \infty$, the only terms that make a significant contribution to the calculation of D_u are those for which $d(x_{ik}, x_{jk})$ is equal to the minimum of all the observed distances for individuals in the same subpopulation. This suggests another statistic,

$$D_\infty = \sum_{k=1}^{L} \sum_{i=1}^{n_k-1} \sum_{j=i+1}^{n_k} 1_{\{d(x_{ik}, x_{jk}) = d_{\min}\}}, \qquad (6)$$

where $d_{\min}$ is the minimum value of $d(x_{ik}, x_{jk})$ over all k and $i \neq j$. Thus, D_∞ is equal to the number of pairs of individuals in the same subpopulation which are separated by the minimum observed distance $d_{\min}$. This statistic can be written in the general form (1) by setting $w_k = 1$ and $F(x) = 1_{\{x = d_{\min}\}}$. In this case the function $F(x)$ depends on the data set being used.

The D_∞ statistic is similar to the S_{nn} statistic of Hudson (2000). S_{nn} measures the frequency at which the individuals most similar to a given individual are found in the same subpopulation as the given individual. Let T_i equal the number of nearest neighbors of individual i, where the nearest neighbors are defined to be those individuals separated from individual i by the minimum observed distance. Also, let W_i equal the number of nearest neighbors of individual i which are in the same subpopulation as individual i. Hudson defines the S_{nn} statistic to be

$$S_{\text{nn}} = \frac{1}{N} \sum_{i=1}^{N} \frac{W_i}{T_i}, \qquad (7)$$

where again $N = \sum n_k$ is the total number of individuals sampled. S_{nn} differs from the D_∞ statistic in that each individual contributes to the calculation of the statistic regardless of the relative distances to the nearest neighbors. For the D_∞ statistic, only those individuals which have the closest nearest neighbors contribute to the calculation.

We predict that the D_u statistics will be asymptotically normal. To explain the reason for the normal distribution we note that (i) under a random labeling of individuals $x(s)$,

$$D_u = \sum_{\{s,\, t\}} F(d(s, t)) \cdot 1_{(x(s) = x(t))},$$

where the sum is over pairs of distinct individuals, and (ii) the random variables $\xi_{s,\,t} = 1_{(x(s) = x(t))}$ which are 1 if the labels of s and t agree and 0 otherwise are almost independent. For example, if we have K groups of a size L for a total population size of N, calculations with the multinomial distribution show that for distinct individuals s, t, u, v,

$$\begin{aligned}
\operatorname{Cov}(\xi_{tu}, \xi_{uv}) &= -\frac{L-1}{N-1} \cdot \left(\frac{N-L}{(N-1)(N-2)} \right) \\
&\geqslant -\frac{1}{N-2} \\
\operatorname{Cov}(\xi_{st}, \xi_{uv}) &= \frac{L-1}{N-1} \cdot \left(\frac{2(N-L)}{(N-1)(N-2)(N-3)} \right) \\
&\leqslant \frac{2}{(N-2)(N-3)}.
\end{aligned}$$

This shows that D_u is a sum of weakly dependent random variables and suggests that the Central Limit Theorem should apply.

However, since the normal distribution is an approximation we will use the permutation-based method outlined by Hudson *et al.* (1992) to derive critical values for our tests. This involves randomly assigning the individuals in the data set to locations, keeping the number of individuals present in each location constant. This procedure is repeated many times and the statistics calculated from the random permutations are compared with the value of the statistic for the data.

POWER ANALYSIS

To determine the power of our D_u statistics for $1 \leqslant u \leqslant \infty$, we constructed sample populations under neutral Wright–Fisher island models and stepping stone models. For the Wright–Fisher island model the coalescent process described by Hudson (1990) was used to construct sequences. We used some of the same parameter values as Hudson *et al.* (1992) and Hudson (2000) so that we could compare our results with theirs. The parameters involved in these simulations were N = the population size, m = the fraction of migrants in

each subpopulation in each generation, and $\mu =$ the neutral mutation rate per generation. The infinite-sites model, which assumes that all mutations occur at new sites, was used. Subpopulations consisting of N diploid individuals were used and there was no recombination.

For the stepping stone model we considered three different spatial neighborhoods: the four nearest neighbors, a 5×5 neighborhood centered at the point, and uniform dispersal across the entire grid (mean field). The parameters involved in these simulations were the grid size, the number of individuals per site, and the number of mutations.

The power of the D_u tests to detect genetic differentiation was calculated by generating 4000 sample populations using each of the two models and different sets of parameters. For each of the samples, D_u was calculated with $u = 1$, 1.01, 1.1, 2, 10, 100, and ∞. The significance of the statistics was calculated using 4000 permutations of the locations so that our setup would be identical to that of Hudson (2000). The power of the test was estimated as the percentage of the sample populations for which the null hypothesis was rejected. We rejected the null hypothesis when $P \leqslant 0.05$.

The power of these new statistics was compared to the power of the sequence-based statistics K_s^* (using the more powerful weighting scheme) and S_{nn} defined earlier as well as to the traditional χ^2 test statistic (Nei, 1987, p. 227):

$$\chi^2 = \sum_{i=1}^{L} \sum_{j=1}^{K} \frac{(n_{ij} - n_i \hat{p}_j)^2}{n_i \hat{p}_j}. \tag{8}$$

Here, L is the number of locations, K is the number of different sequences or haplotypes in the entire sample, n_i is the sample size from location i, n_{ij} is the number of copies of sequence j observed in location i, $\hat{p}_j = (1/N) \sum_{i=1}^{L} n_{ij}$ is the frequency of sequence j in the sample, and $N = \sum_{i=1}^{L} n_i$ is the total number of individuals sampled. To allow for the presence of rare haplotypes, we computed the significance of the χ^2 statistic directly using a permutation approach, rather than using a limit theorem to assert that this has approximately a χ^2 distribution with $(L-1)(K-1)$ degrees of freedom. The permutation approach allowed us to get the most information out of the χ^2 test since lumping of rare haplotypes can result in a loss of power of the test (Hudson, 1992; Roff and Bentzen, 1989, 1992).

Power estimations using the Wright–Fisher island model were made for three different sets of population parameters and five different sample sizes of two subpopulations (Table I). Looking at the D_u statistic for different values of u, we saw that the power increased as u increased, reaching a maximum at $u = 2$ or 10. The power decreased for values larger than these. The most powerful D_u test was more powerful than the K_s^* test in all of the cases examined. Comparing the D_u test and the χ^2 test, we see that the latter was more powerful in the majority of cases. The D_u test was often more powerful than the χ^2 test when the sample size was small. Finally, we see that the S_{nn} statistic was the most powerful in all cases examined. However, the difference between the maximum power of the D_u test and the power of the S_{nn} test was quite small for some of the simulations and less than 0.11 for all of them.

TABLE I

Fraction of 4000 Simulations (Wright–Fisher Island Model) Rejecting the Null Hypothesis

$4Nm$[a]	2.0	2.0	2.0	5.0	2.0
$4N\mu$	5.0	5.0	5.0	5.0	1.0
(n_1, n_2)	(10, 10)	(35, 5)	(25, 25)	(25, 25)	(25, 25)
K_s^*	0.626	0.602	0.935	0.684	0.766
D_1	0.512	0.403	0.814	0.520	0.708
$D_{1.01}$	0.529	0.413	0.824	0.532	0.724
$D_{1.1}$	0.577	0.489	0.895	0.610	0.740
D_2	0.701	0.717	0.975	0.807	0.779
D_{10}	0.666	0.724	0.966	0.817	0.770
D_{100}	0.664	0.718	0.963	0.811	0.766
D_∞	0.542	0.683	0.954	0.792	0.756
S_{nn}	0.733	0.779	0.996	0.925	0.871
χ^2	0.600	0.793	0.993	0.909	0.857

[a] $N =$ population size, $m =$ migration rate, $\mu =$ mutation rate, $n_i =$ sample size for population i.

Power estimations using the stepping stone model were made for six different cases (Table II). All of the simulations were completed on a grid of size 25×25 with 50 individuals per site. In each case all individuals from two of the subpopulations were used in the calculation of the statistics. The spatial neighborhood, number of mutations, and the distance between sampled subpopulations varied between the simulations. Again we saw that the power of the D_u statistic increased as u increased, reaching a maximum for some $u \geqslant 10$. When the number of mutations was small ($\leqslant 50$), the power decreased as u approached infinity. The power of the K_s^* test appeared to lie either between that of $D_{1.1}$ and D_2 or between D_2 and D_{10} and was less than D_{10} in all cases studied. In all of the cases examined we found that the χ^2 test was more powerful than the D_u statistics. Again, the S_{nn} statistic was the

TABLE II

Fraction of 4000 Simulations[a] (Stepping Stone Model) Rejecting the Null Hypothesis

Spatial neighborhood[b]	mf	5×5	4	4	4	4
Number of mutations	100	100	100	50	10	10
Compared site (0,0) to	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(9, 9)
K_s^*	0.988	0.970	0.925	0.881	0.654	0.814
D_1	0.948	0.923	0.890	0.854	0.645	0.809
$D_{1.01}$	0.950	0.929	0.900	0.857	0.647	0.812
$D_{1.1}$	0.978	0.956	0.912	0.867	0.649	0.812
D_2	0.992	0.978	0.934	0.884	0.651	0.813
D_{10}	0.992	0.979	0.936	0.889	0.656	0.815
D_{100}	0.992	0.979	0.936	0.889	0.657	0.814
D_∞	0.992	0.979	0.936	0.887	0.655	0.812
S_{nn}	0.999	0.994	0.964	0.927	0.710	0.852
χ^2	0.999	0.993	0.960	0.920	0.676	0.826

[a] All simulations were completed on a 25×25 grid with 50 individuals per site.

[b] mf = mean field, $5\times5=5\times5$ neighborhood, $4=4$ nearest neighbors.

most powerful for all of the cases examined. However, the difference between the maximum power of the D_u test and the power of the S_{nn} test was again quite small and was less than 0.06 for all cases examined.

It is important to note that none of the power calculations considered genes with recombination. Hudson *et al.* (1992) found that without recombination, the test based on χ^2 was more powerful than all of the sequence statistics they examined. However, in many cases with recombination, the sequence statistics (including K_s^*) were found to be more powerful than the haplotype-based χ^2 statistic. This suggests that the D_u test may be more powerful than the χ^2 statistic when using genes with recombination, which is consistent with the fact that our test produces more significant results than the χ^2 test when applied to real data.

APPLICATIONS

Drosophila Microsatellite Data

Our first example is a data set collected by Wetterstrand (1997) on 16 microsatellite loci in 99 isofemale lines of *Drosophila melanogaster* individuals, sampled from five different locations: Australia ($n=20$), China ($n=20$), Ecuador ($n=19$), Zimbabwe ($n=20$), and the United States ($n=20$). Using the proportion of nonshared alleles as a distance, she built a neighbor joining tree. The result, taken from her thesis, is given in Fig. 1. We defined the distance between individuals on her tree as the number of internal nodes crossed on the shortest path connecting the two individuals.

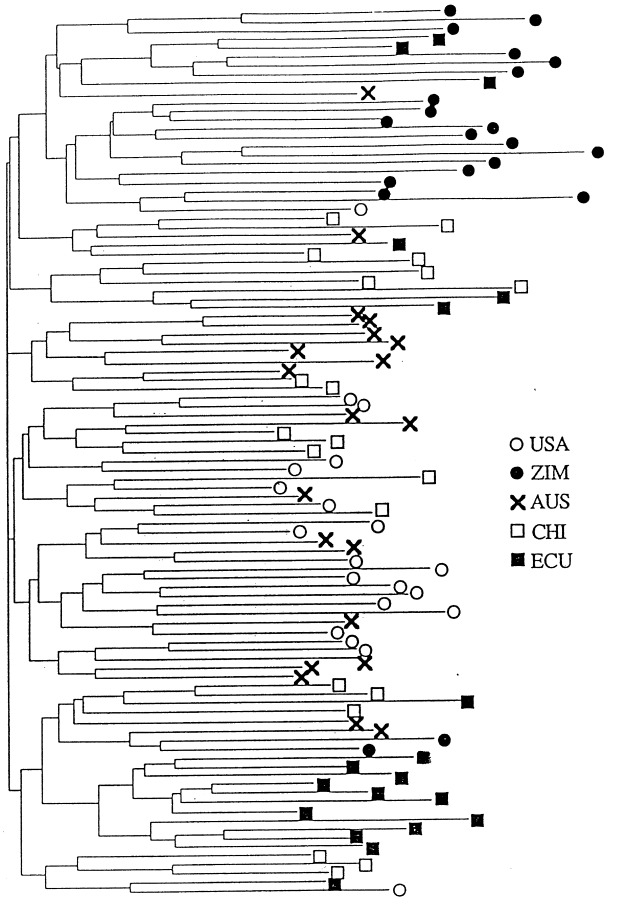

FIG. 1. The neighbor joining tree for the Wetterstrand (1997) microsatellite date for *Drosophila melanogaster*. Individuals were sampled from five locations: Australia, China, Ecuador, the United States, and Zimbabwe. The tree is based on the proportion of shared alleles between individuals and is taken from Wetterstrand's 1997 M.Sc. thesis with permission.

The test statistics for the D_u, K_s^*, and S_{nn} tests were calculated and the significance was determined using 100,000 permutations of the locations of the individuals. All of the tests indicated significant genetic differentiation of the five subpopulations ($P<10^{-5}$). The 100,000 random colorings of the tree for the D_2 statistic gave a distribution that appeared to be close to normal with mean 9.3 and standard deviation 1.4 (see Fig. 2).

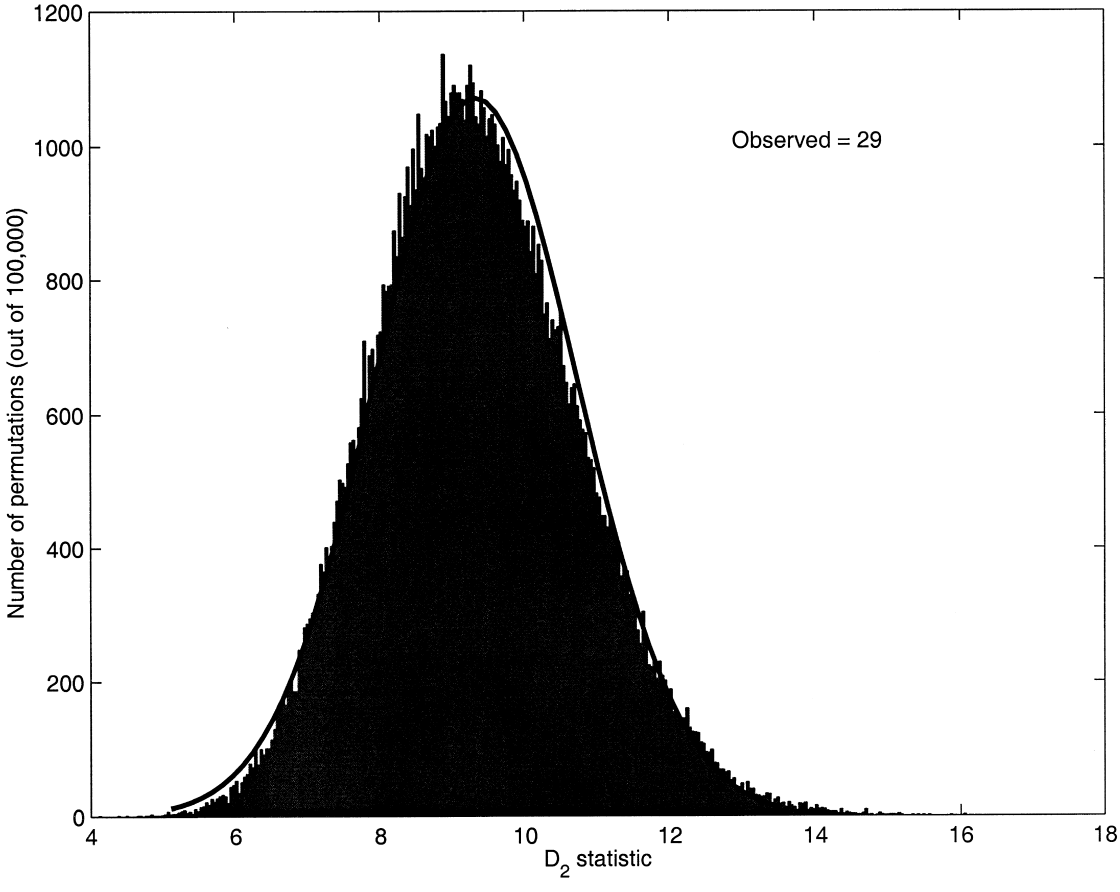

FIG. 2. Histogram of the D_2 statistic for 100,000 random permutations of the locations of individuals on the Wetterstrand (1997) neighbor joining tree. The curve is the normal distribution with mean 9.3 and standard deviation 1.4. The D_2 statistic for the observed tree is 29.

Drosophila DNA Sequence Data

Our second example is a data set of Hamblin and Veuille (1999) containing DNA sequence data from *Drosophila simulans* individuals. Data for a 735 bp region at the *vermilion* locus was collected from 82 *D. simulans* individuals from seven different locations: Cameroon ($n = 12$), Italy ($n = 12$), Kenya ($n = 13$), Lesser Antilles ($n = 12$), Tanzania ($n = 11$), the United States ($n = 12$), and Zimbabwe ($n = 10$). To analyze their data, we defined the distance between two individuals to be the number of pairwise differences between their sequences. We computed all of the statistics for the *vermilion* locus for a variety of population comparisons (Table III). The first comparison was of all seven populations. Since the United States and Lesser Antilles (an arc of islands in the Caribbean Sea) are both separated from the other populations by the Atlantic Ocean, it should not be surprising that the result was highly significant: the statistic for the data was smaller than all 10,000 permutations for the K_s^* and D_1 statistics and larger than all 10,000 permutations for the remaining statistics ($P < 0.0001$).

Reducing our focus to the Italian and four African locations, we again found highly significant results for all statistics ($P < 0.0001$). Dropping the Italian population, the comparison among the four African populations led to a significant result for K_s^*, D_u with $u \leqslant 100$, and S_{nn} ($P < 0.0475$). However, χ^2 and D_∞ gave nonsignificant results ($P > 0.06$). For all of the three-population comparisons we found significant results using the D_u statistic with $u \leqslant 1.1$ and K_s^* ($P < 0.03$). Using the first letter of the name of the country as an abbreviation, we found that the CKZ and CTZ comparisons were also significant using the D_u statistic with $u \leqslant 100$. The S_{nn} and χ^2 statistics were only significant for the CKZ comparison. The D_u and K_s^* statistics therefore suggest that none of the possible three-way groupings of African populations can be considered homogeneously mixing at the *vermilion*

TABLE III

Results of the Tests for Population Subdivision

Statistic	K_s^*	D_1	$D_{1.01}$	$D_{1.1}$	D_2	D_{10}	D_{100}	D_∞	S_{nn}	χ^2
Vermilion										
All	**0.0001**[b]	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**
AI[a]	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**
A	**0.0012**	**0.0002**	**0.0002**	**0.0009**	**0.0108**	**0.0475**	**0.0475**	0.1055	**0.0142**	0.0611
CKZ	**0.0003**	**0.0002**	**0.0002**	**0.0003**	**0.0110**	**0.0486**	**0.0486**	0.1134	**0.0006**	**0.0498**
CTZ	**0.0046**	**0.0016**	**0.0017**	**0.0031**	**0.0134**	**0.0343**	**0.0343**	0.0829	0.0567	0.0539
KTZ	**0.0104**	**0.0033**	**0.0036**	**0.0077**	0.0626	0.0718	0.0718	0.1247	0.0964	0.0733
CKT	**0.0266**	**0.0162**	**0.0162**	**0.0193**	0.0916	0.3090	0.3090	0.5897	0.2391	0.4129
KZ	**0.0012**	**0.0008**	**0.0008**	**0.0013**	**0.0248**	**0.0323**	**0.0323**	0.1090	0.0877	0.1106
CK	**0.0067**	**0.0037**	**0.0037**	**0.0058**	0.0523	0.2015	0.2015	0.5001	**0.0093**	0.4478
TZ	**0.0186**	**0.0115**	**0.0121**	**0.0226**	0.0628	0.0644	0.0742	0.1043	0.3678	0.1043
CZ	**0.0222**	**0.0116**	**0.0116**	**0.0122**	0.0652	0.1212	0.1212	0.2627	0.0561	0.2248
CT	**0.0391**	**0.0259**	**0.0261**	**0.0330**	**0.0360**	0.1263	0.1265	0.6081	0.3191	0.4635
KT	0.8545	0.8605	0.8467	0.8006	0.7672	0.7654	0.8064	1.0000	0 .7873	1.000
G6pd										
AE	**0.0003**	**0.0004**	**0.0004**	**0.0004**	**0.0001**	**0.0156**	**0.0225**	**0.0390**	**0.0002**	**0.0100**
A	**0.0058**	**0.0064**	**0.0057**	**0.0047**	**0.0181**	0.2568	0.2998	0.4349	**0.0327**	0.2579
CKZ	**0.0004**	**0.0005**	**0.0005**	**0.0004**	**0.0008**	0.1092	0.1685	0.2954	**0.0093**	0.1976
CTZ	**0.0055**	**0.0120**	**0.0110**	**0.0081**	**0.0196**	0.2002	0.2122	0.3550	**0.0121**	0.1661
CKT	**0.0273**	**0.0264**	**0.0232**	**0.0236**	0.0636	0.3371	0.3556	0.5130	0.0811	0.3972
KTZ	0.2917	0.2721	0.2524	0.2328	0.3827	0.5787	0.5810	0.6634	0.3626	0.4024
CZ	**0.0009**	**0.0022**	**0.0020**	**0.0014**	**0.0008**	**0.0214**	**0.0262**	0.2058	**0.0051**	0.1624
CK	**0.0017**	**0.0023**	**0.0020**	**0.0019**	**0.0022**	0.0658	0.0847	0.2126	**0.0049**	0.2029
TZ	0.1105	0.1717	0.1526	0.1289	0.1876	0.3216	0.3219	0.4065	0.1319	0.2603
CT	0.2333	0.3701	0.3187	0.3155	0.3611	0.5461	0.5461	0.7753	0.1857	0.5960
KZ	0.3951	0.2818	0.2629	0.2562	0.4058	0.6347	0.6347	0.8441	0.6098	0.7856
KT	0.5660	0.5491	0.5165	0.5186	0.6357	0.6550	0.6550	0.6831	0.6123	0.5980
Walleye										
Spacer	0.1128	0.8468	0.8437	0.8279	0.7273	0.6380	0.6278	0.6309	0.3507	0.3543
Control	0.0724	**0.0289**	**0.0286**	**0.0278**	**0.0244**	**0.0240**	**0.0244**	**0.0246**	0.0996	**0.0384**
Both	0.0717	0.1582	0.1553	0.1438	0.0693	**0.0371**	**0.0259**	**0.0271**	**0.0240**	**0.0242**

[a] A = Africa, C = Cameroon, E = Europe, I = Italy, K = Kenya, T = Tanzania, Z = Zimbabwe.
[b] Estimated P value based on 10,000 random partitions.

locus. Both the S_{nn} and the χ^2 statistics fail to recognize this and suggest that only the grouping of Cameroon, Kenya, and Zimbabwe does not form a homogeneously mixing population.

Looking at the pairwise comparisons of the African populations we again see a discrepancy between the predictions of the statistics. For all of the pairwise comparisons except for that of Kenya with Tanzania, the K_s^* and D_u statistics with $u \leqslant 1.1$ are significant. However, the S_{nn} statistic is only significant for the CK comparison ($P \leqslant 0.01$) and the χ^2 statistic is not significant for any of the comparisons ($P \geqslant 0.1$).

Hamblin and Veuille (1999) also considered data for a 700-bp region of the third exon of the *G6pd* locus from 66 *D. simulans* individuals from six different locations: Cameroon ($n = 12$), Europe (Italy and France) ($n = 9$), Kenya ($n = 12$), Lesser Antilles ($n = 12$), Tanzania ($n = 10$), and Zimbabwe ($n = 11$). As before, consideration of all six populations or just the five in Africa and Europe led to very significant results for all statistics, so we concentrated on the four African populations. For the four African population comparison and the CKZ and CTZ three-population comparisons (see Table III) we found significant results for K_s^*, D_u with $u \leqslant 2$, and S_{nn} ($P \leqslant 0.04$). The CKT comparison was also significant for K_s^* and D_u with $u \leqslant 1.1$ ($P \leqslant 0.03$). The remaining three-way comparison, KTZ, was not significant for any statistic. This suggests the possibility that the three populations are similar enough genetically to be considered a homogeneously

mixing unit at the *G6pd* locus. Further evidence of this comes from the nonsignificance of all of the statistics calculated for the pairwise comparisons of these three populations (Table III).

We again see a discrepancy between the conclusions of the statistics for the pairwise comparisons of the African populations. While the K_s^*, D_u with $u \leqslant 2$, and S_{nn} statistics are significant for the comparisons of *CK* and *CZ* ($P \leqslant 0.03$), the χ^2 statistic does not identify significant differentiation for any pair of populations.

Hamblin and Veuille (1999) studied the genetic differentiation between populations using F_{ST} estimates for all pairwise comparisons. Because the pairwise tests were nonindependent, they analyzed their results qualitatively rather than choosing a significance threshold. For both loci we reached the conclusion of Hamblin and Veuille (1999) that Kenya and Tanzania are essentially homogeneously mixing. For the *vermilion* locus, Hamblin and Veuille identified three genetically distinct groups in Africa: Zimbabwe, Cameroon, and Tanzania/Kenya. While the K_s^* and D_u test with $u \leqslant 1.1$ also identified these groups, the S_{nn} and χ^2 tests failed to reach this conclusion.

Using the *G6pd* locus, Hamblin and Veuille concluded that while Cameroon is differentiated from all populations except Tanzania, the remaining populations show no significant differentiation from one another. This conclusion is exactly that reached by the K_s^*, D_u with $u \leqslant 2$, and S_{nn} tests. The χ^2 test fails to recognize this differentiation.

It is interesting to note that while the P values of the D_u tests for the comparisons at the *vermilion* locus decreased as u decreased, reaching a minimum P value as u approached 1, the P values for the comparisons at the *G6pd* locus reached a minimum when $u = 1.01$, 1.1, or 2. In all of the above comparisons, the minimum P value of the D_u statistics was much smaller than the P values of the K_s^*, S_{nn}, and χ^2 statistics, giving highly significant results for some comparisons that were missed by both the S_{nn} and the χ^2 tests.

We also note that for many of the comparisons, the P values for the K_s^* test fell in between the P values of the $D_{1.1}$ and D_2 tests. This result is analogous to that found in the power calculations for the stepping stone model.

Walleye Pollock Mitochondrial DNA Sequences

Our final example is a data set of Shields and Gust (1995). They examined the 76-bp spacer region from mitochondrial DNA sequences of 110 walleye pollock, *Theragra chalcogramma*, individuals divided into the following regions: Southwest Bering Sea ($n = 18$), Gulf of Alaska ($n = 9$), North Bering Sea ($n = 12$), Donut Hole ($n = 7$), West Aleutians ($n = 17$) and East Aleutians ($n = 47$). Twenty unique haplotypes, based on 20 mutations at 16 segregating sites, were identified, with 83 individuals having the same haplotype. They performed a χ^2 analysis on the data by lumping the rare haplotypes together. This method indicated no significant genetic differentiation of the subpopulations and had $P < 0.282$. Likewise, all of the statistics discussed previously indicated no significant differentiation ($P \geqslant 0.1$) (Table III).

Shields and Gust (1995) also investigated a 250-bp region of the control region sequence from 140 individuals divided into the same regions as before, but with slightly different sample sizes: Southwest Bering ($n = 20$), Gulf of Alaska ($n = 8$), North Bering ($n = 18$), Donut Hole ($n = 8$), West Aleutians ($n = 23$), and East Aleutians ($n = 63$). Seventeen unique haplotypes with 14 segregating sites were identified, with 114 individuals having the same haplotype. Again, they found no significant differentiation using a χ^2 analysis with rare haplotypes lumped together ($P \leqslant 0.212$). The K_s^* and S_{nn} tests were also nonsignificant ($P \geqslant 0.07$). However, the permutation-based χ^2 test and the D_u test for all u were significant ($P \leqslant 0.04$). The minimum value for the D_u test occurred when $u = 10$.

Shields and Gust (1995) made one more comparison by looking at the spacer and control regions together. They identified 21 unique haplotypes with 19 segregating sites in the sample. In order to find some geographic variability, they used four of the original regions to create two new regions by combining the Southwest Bering and Northern Bering samples into a Western Bering sample and combining the Western and Eastern Aleutian samples into a Eastern Bering sample. This new comparison contained 80 individuals: Western Bering ($n = 32$) and Eastern Bering ($n = 48$). Using a χ^2 test on this comparison, Shields and Gust (1995) found significant genetic differentiation with $P \leqslant 0.021$. The S_{nn} and χ^2 tests gave similar results with $P \leqslant 0.025$. The D_u test was significant for $u \geqslant 10$ and reached a minimum value of $P = 0.0259$ when $u = 100$. The K_s^* test was again nonsignificant ($P \geqslant 0.07$).

DISCUSSION

We have developed a family of sequence statistics, D_u, $1 \leqslant u \leqslant \infty$ that can be used to test for population subdivision. These statistics measure the overall amount

of variation within subpopulations by summing an exponential function of the distance between individuals from the same subpopulation. We examined the power of these new statistics to detect differentiation under both a neutral Wright–Fisher island model and a stepping stone model with three definitions of spatial neighborhoods. The power of the new D_u statistics was compared to that of the K_s^*, S_{nn}, and χ^2 tests. Simulation results indicated that while the S_{nn} statistic was more powerful under all conditions examined, the magnitude of the differences between the powers of the S_{nn}, χ^2, and D_u statistics was quite small in most cases.

Using the statistics on various data sets we found that the D_u, S_{nn}, and χ^2 statistics perform very differently. Consideration of *Drosophila* sequence data from the *vermilion* and *G6pd* loci showed that in some circumstances, one gets much more significant results using D_u in the limit as as $u \to 1$. Data from the walleye pollock mitochondrial DNA gave the opposite result and one gets the most significant result using D_u in the limit as $u \to \infty$. One possible reason for this difference is that while the two *Drosophila* data sets are from nuclear loci, the walleye pollock data is mitochondrial so there is no recombination. The two *Drosophila* loci are believed to undergo high rates of recombination and have scaled recombination rates of $Nr/\text{base pair} = 0.01$ for *vermilion* and $Nr/\text{base pair} = 0.005$ for *G6pd* (Hamblin and Veuille, 1999). This gives the estimates for the whole region surveyed of $4Nr = 29.4$ for *vermilion* and $4Nr = 14$ for *G6pd*. For genes with high rates of recombination, Hudson *et al.* (1992) predict that the sequence statistic K_s^* will be more powerful than the haplotype statistic χ^2. Our results showed that the P value of the K_s^* statistic often falls between that of $D_{1.1}$ and D_2. Therefore, we predict that the D_u statistics should be at least as powerful as the K_s^* statistic and in many cases much more powerful than both the K_s^* and χ^2 tests. For the walleye pollock data sets there is no recombination and the heterozygosity is much smaller than that of the *D. simulans* data. In this case the D_u statistics with $u \geqslant 10$ produce results equivalent to the χ^2 test and more significant than the K_s^* test.

The presence or lack of recombination does not explain the different performances of the S_{nn} and D_u statistics. Hudson (2000) found that in a large number of simulations, including those with and without recombination, the S_{nn} statistic was more powerful than both the K_s^* and the χ^2 statistic. The power tests included here for both the island and the stepping stone models agree with this conclusion and produce the additional result that S_{nn} is more powerful than all of the D_u statistics.

However, we see that when using the statistics on real data, the D_u tests frequently identify genetic differentiation that is missed by the S_{nn} tests. The causes of these differences are unknown and merit further study.

We do not know how to predict *a priori* which of our D_u statistics will be the best to use on a given data set. However, from the data sets we have analyzed it seems one can afford to try three forms: D_1, D_2, and D_∞, even though one must multiply the P values by 3 to account for the Bonferroni correction.

ACKNOWLEDGMENTS

We thank Linda Buttel for assistance with the spatial simulations, Chip Aquadro for many helpful suggestions, and Dick Hudson for a very thorough referee's report. He reran most of the simulations and uncovered some bugs in our code. Durrett's research was partially supported by grants from the probability program at the National Science Foundation and a supplement to Aquadro's NIH grant.

REFERENCES

Begun, D. J., and Aquadro, C. F. 1995. Molecular variation at the *vermilion* locus in geographically diverse populations of *Drosophila melanogaster* and *D. simulans*, *Genetics* **140**, 1019–1032.

Charlesworth, B. 1998. Measures of divergence between populations and the effect of forces that reduce variability, *Mol. Biol. Evol.* **15**, 538–543.

Charlesworth, B., Nordborg, M., and Charlesworth, D. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations, *Genet. Res. Camb.* **70**, 155–174.

Eanes, W. F., Kirchner, M., and Yoon, J. 1993. Evidence for adaptive evolution of the *G6pd* gene in the *Drosophila melanogaster* and *Drosophila simulans* lineages, *Proc. Natl. Acad. Sci. USA* **90**, 7475–7479.

Eanes, W. F., Kirchner, M., Yoon, J., Biermann, C. H., Wang, I., McCartney, M. A., and Verrelli, B. C. 1996. Historical selection, amino acid polymorphism and lineage-specific divergence at the *G6pd* locus in *Drosophila melanogaster* and *D. simulans*, *Genetics* **144**, 1027–1041.

Excoffier, L. P., Smouse, E., and Quattro, J. M. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data, *Genetics* **131**, 479–491.

Goudet, J., Raymond, M., De Meeüs, T., and Rousset, F. 1996. Testing differentiation in diploid populations, *Genetics* **144**, 1933–1940.

Hamblin, M. T., and Veuille, M. 1999. Population structure among African and derived populations of *D. simulans*: Evidence for ancient subdivision and recent admixture, *Genetics* **153**, 305–317.

Holsinger, K. E., and Mason-Gamer, R. J. 1996. Hierarchical analysis of nucleotide diversity in geographically structured populations, *Genetics* **142**, 629–639.

Hudson, R. R. 1990. Gene genealogies and the coalescent process, *in* "Oxford Surveys in Evolutionary Biology" (D. J. Futuyma and J. Antonovics, Eds.), Vol. 7, pp. 1–44, Oxford Univ. Press, London.

Hudson, R. R. 1992. Reply to Roff and Bentzen, *Mot. Biol. Evol.* **9**, 969.

Hudson, R. R. 2000. A new statistic for detecting genetic differentiation, *Genetics* **155**, 2011–2014.

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

Lynch, M., and Crease, T. J. 1990. The analysis of population survey data on DNA sequence variation, *Mol. Biol. Evol.* **7**, 377–394.

Michalakis, Y., and Excoffier, L. 1996. A generic estimation of population subdivision using distances between alleles with special references for microsatellite loci, *Genetics* **142**, 1061–1064.

Nei, M. 1973. Analysis of gene diversity in subdivided populations, *Proc. Natl. Acad. Sci. USA* **70**, 3321–3323.

Nei, M. 1982. Evolution of human races at the gene level, *in* "Human Genetics, Part A: The Unfolding Genome" (B. Bonne-Tamir, T. Cohen, and R. M. Goodman, Eds.), pp. 167–181, A. R. Liss, New York.

Nei, M. 1987. "Molecular Evolutionary Genetics," Columbia Univ. Press, New York.

Neigel, J. E. 1997. A comparison of alternative strategies for estimating gene flow from genetic markers, *Annu. Rev. Ecol. Syst.* **28**, 105–128.

Petit, R. J., and Pons, O. 1998. Bootstrap variance of diversity and differentiation estimators in a subdivided population, *Heredity* **80**, 56–61.

Raymond, M., and Rousset, F. 1995. An exact test for population differentiation, *Evolution* **49**, 1280–1283.

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

Roff, D. A., and Bentzen, P. 1992. Detecting geographic subdivision: A comment on a paper by Hudson *et al.*, *Mol. Biol. Evol.* **9**, 968.

Rousset, F. 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance, *Genetics* **145**, 1219–1228.

Shields, G. F., and Gust, J. R. 1995. Lack of geographic structure in mitochondrial DNA sequences of Bering Sea walleye pollock, *Theragra chalcogramma, Mol. Marine Biol. Biotechnol.* **4**, 69–82.

Slatkin, M. 1995. A measure of population subdivision based on microsatellite allele frequencies, *Genetics* **139**, 457–462.

Slatkin, M., and Maddison, W. P. 1989. A cladistic measure of gene flow inferred from the phylogenies of alleles, *Genetics* **123**, 603–613.

Takahata, N., and Nei, M. 1984. Letter to the Editor: F_{ST} and G_{ST} statistics in the finite island model, *Genetics* **107**, 501–504.

Templeton, A. R. 1998. Nested clade analyses of phylogeographic data: Testing hypotheses about gene flow and population history, *Mol. Ecol.* **7**, 381–397.

Turner, T. F., Trexler, J. C., Harris, J. L., and Haynes, J. L. 2000. Nested cladistic analysis indicates population fragmentation shapes genetic diversity in a freshwater mussel, *Genetics* **154**, 777–785.

Weir, B. S. 1983. "Statistical Analysis of DNA Sequence Data," Dekker, New York.

Weir, B. S. 1996. "Genetic Data Analysis II," Sinauer, Sunderland, MA.

Wetterstrand, K. S. 1997. "Microsatellite Polymorphism and Divergence in Worldwide Populations of *Drosophila Melanogaster* and *D. simulans*," M.Sc. thesis, Cornell University, Ithaca, NY.

Wright, S. 1951. The genetical structure of populations, *Ann. Eugen.* **15**, 323–354.