A Test of Neutrality Based on Interlocus Associations

John K. Kelly[1]

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

Manuscript received October 21, 1996
Accepted for publication April 18, 1997

ABSTRACT

The evolutionary processes governing variability within genomic regions of low recombination have been the focus of many studies. Here, I investigate the statistical properties of a measure of intrlocus genetic associations under the assumption that mutations are selectively neutral and sites are completely linked. This measure, denoted Z_{nS}, is based on the squared correlation of allelic identity at pairs of polymorphic sites. Upper bounds for Z_{nS} are determined by simulations. Various deviations from the neutral model, including several different forms of natural selection, will inflate the value of Z_{nS} relative to its neutral theory expectations. Larger than expected values of Z_{nS} are observed in genetic samples from the *yellow-ac-scute* and *Adh* regions of *Drosophila melanogaster.*

THERE is now a substantial body of statistical theory to test hypotheses regarding gene sequence evolution (EWENS 1990; HUDSON 1990; KREITMAN 1990). Much of this theory predicts the patterns of genetic variation that are likely to arise in the absence of natural selection (KIMURA 1983). These neutral models provide a null hypothesis that can be tested against genetic data. They also provide a basis for comparison when models involving natural selection are considered.

The simplest tests of neutrality concern gene sequence variation within samples from a single population at a single genetic locus (a nonrecombining sequence). The current set of test statistics in this category measure the frequency distribution of mutant alleles within the sample (WATTERSON 1978; TAJIMA 1989a; FU and LI 1993; BRAVERMAN *et al.* 1995; SIMONSEN *et al.* 1995). A second important characteristic of gene sequence variation, not directly measured by these tests, is the pattern of associations among mutant alleles at different polymorphic sites.

The most commonly used measure of interlocus genetic associations is the linkage disequilibrium. Consider a population of sequences that is polymorphic for two alternative alleles at a series of nucleotide sites. Let p_i and p_j denote the frequency of the mutant allele at the ith and jth loci, respectively. The linkage disequilibrium between loci i and j, denoted D_{ij}, is

$$D_{ij} = p_{ij} - p_i p_j, \qquad (1)$$

where p_{ij} is the frequency of sequences that have mutant alleles at both sites.

The magnitude of the linkage disequilibrium depends on both the strength of association between the two loci and the frequency of mutant alleles at each locus. A standardized measure of linkage disequilibrium (ranging from 0 to 1) is δij, the squared correlation of allelic identity between loci i and j (HARTL and CLARK 1989, pp. 53–54):

$$\delta_{ij} = \frac{D_{ij}^2}{\mathrm{p}_i(1 - p_i)p_j(1 - p_j)}. \qquad (2)$$

It is noteworthy that δ_{ij} yields the same value regardless of which alternative allele at each locus is considered the mutant. Thus, it can be calculated without information about the ancestral allele at each site.

In general usage, the linkage disequilibrium (and any statistic derived from it) is defined by the haplotype frequencies within the entire population. Here, we will be concerned with the values of D_{ij} and δ_{ij} within a sample of gene sequences (calcuated from the haplotypic frequencies within the sample). I investigate the properties of a sample statistic, Z_{nS}, that averages δ_{ij} over all pairwise comparisons of S polymorphic sites in a sample of n sequences:

$$Z_{nS} = \frac{2}{S(S-1)} \sum_{i=1}^{S-1} \sum_{j=i+1}^{S} \delta_{ij}. \qquad (3)$$

The probability distribution of $Z_{n\mathrm{S}}$, conditional on n and S, is investigated via computer simulations. The simulations are limited to neutral evolution within sequences of completely linked sites. The consequences of various deviations from the neutral model on Z_{nS} are discussed with specific attention to the effects of selection at linked sites. Finally, Z_{nS} is estimated from previously published genetic data of *Drosophila melanogaster* from the *yellow-ac-scute* region (MARTIN-CAMPOS *et al.* 1992) and the *Adh* locus (KREITMAN 1983; LAURIE *et al.* 1991).

SAMPLE GENEALOGIES

The expected patterns of genetic variation within a set of gene sequences from a natural population can

[1] *Present address:* John Kelly, Department of Biology, University of Oregon, Eugene, OR 97403-5219.
E-mail: jkk@darkwing.uoregon.edu

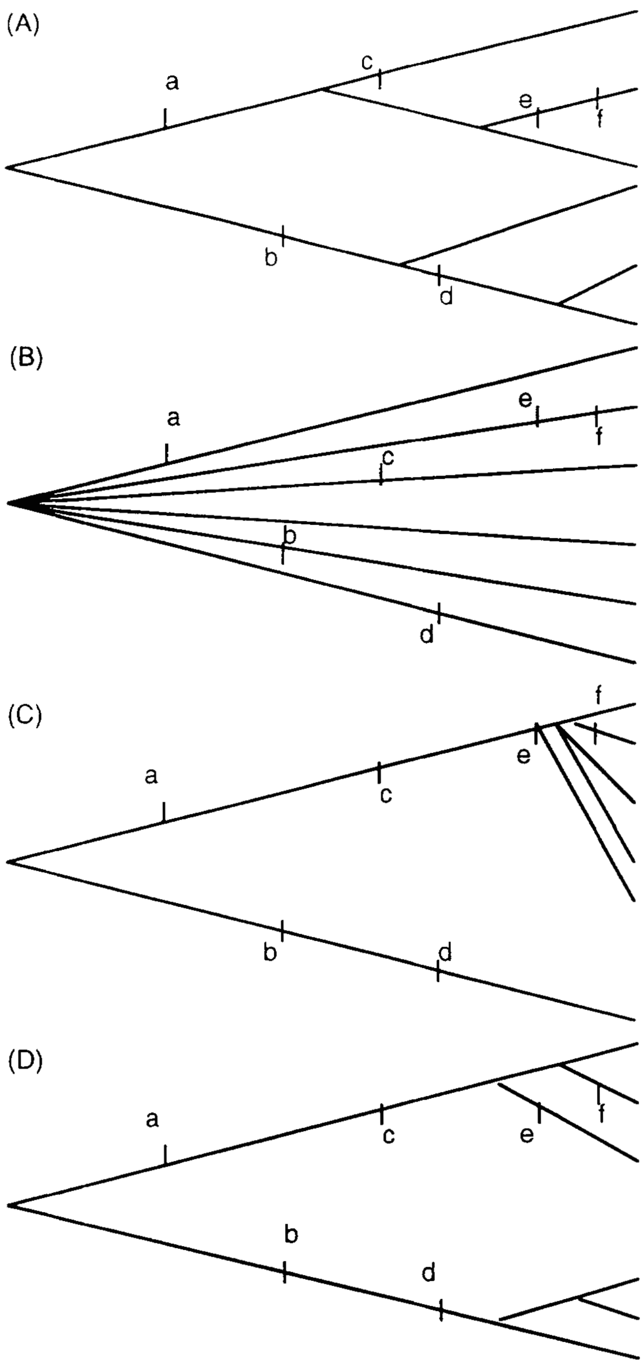

FIGURE 1.—Hypothetical sample genealogies under the neutral model (A), a recent selective sweep at a completely linked site (B), a recent selective sweep at a closely linked site (C), and balancing selection at a closely linked site (D). Mutations occur on the genealogies at points denoted by the lowercase letters a–f. The calculated values of Z_{nS} are as follows: (A) 0.31, (B) 0.10, (C) 0.68, and (D) 0.51.

be investigated by considering the "sample genealogy." When there is no recombination within a sequence, each sequence within a sample has a single ancestral lineage. Figure 1A is a typical sample genealogy under a neutral model of molecular evolution. Viewing time retrospectively, from the present to the past, distinct lineages will "coalesce" at points of shared ancestry. Eventually, all sequences lineages will converge on a single common ancestor (KINGMAN 1982).

This sample genealogy implies certain relationships between sequences and constrains the patterns of genetic variation that may be observed in that sample. There is an extensive mathematical theory describing the stochastic properties of sample genealogies for sequences evolving via neutral mutations (WATTERSON 1975; KINGMAN 1982; TAVARE 1984). I will mention two important features of this theory for the special case of a population governed by a Wright-Fisher demographic model (random mating, diploidy, population size constancy, discrete non-overlapping generations, and binomially distributed individual reproductive success) and an infinite sites mutation model (all mutations occur at previously monomorphic sites and the number of mutations introduced during production of a progeny sequence is Poisson distributed).

Let T_k denote the total time (measured in units of $2N$ generations) during which the genealogy has exactly k lineages, where k ranges from 2 to n. Under the neutral model described above, the T_k are exponentially distributed and stochastically independent. The total branch length of the genealogy, denoted T_{cum}, is equal to

$$T_{cum} = \sum_{j=2}^{n} j\, T_j. \tag{4}$$

Second, the mutational process is stochastically independent of the genealogical process. In figurative terms, mutations are randomly "sprinkled" onto the sample genealogy and each mutation is present in all descendants of the sequence onto which it is dropped (HUDSON 1990). The total number of mutations on the genealogy, which equals the total number of polymorphic sites in the sample, is a Poisson random variable with the mean equal to the product of the mutation rate and T_{cum}.

Effect of population genealogy on δij: In the absence of recombination, δij is a measure of allele frequency equivalency across loci and will equal 1 only if two of the four possible two-locus haplotypes are present in a sample ($p_i = p_j = p_{ij}$). It is less than 1 if three haplotypes are present. Imagine that mutations occur on the genealogy in Figure 1A at the points denoted a, b, c, d, e, and f. Fifteen contrasts between polymorphic sites are averaged to determine Z_{nS}. Two of these contrasts, δab and δef, are equal to 1. The remaining 13 contrasts yield δ_{ij} that are less than 1.

All contrasts between polymorphic sites occurring on lineages that go back in time, unbifurcated, directly to the common ancestor of the entire sample, give δ_{ij} values of 1. This "critical region" of the genealogy includes both lineages during the T_2 time interval and some subsequent segments. For this reason, we expect larger values of Z_{nS} for genealogies with a long period of history with only two ancestors (if T_2 is large, the critical region should constitute a high proportion of the genealogy). The time T_2 has the largest expectation and the largest variance of all time intervals under the coalescent model (HUDSON 1990). Thus, we expect random variation in T_2 to generate much of the variation in Z_{nS}.

Deviations from the neutral genealogy can arise from a number of factors and theoretical studies have explored how different evolutionary processes, including selection at linked sites, affect sample genealogies (HUDSON and KAPLAN 1988; KAPLAN *et al.* 1988, 1989; TAKAHATA 1988; SLATKIN 1989; TAJIMA 1989b). Figure 1B illustrates a sample genealogy for a neutral sequence if it is completely linked to a site where an advantageous mutation has recently swept to fixation (MAYNARD SMITH and HAIGH 1974; KAPLAN *et al.* 1989; BRAVERMAN *et al.* 1995). The same type of genealogy may result if the population has recently experienced a severe bottleneck in size (TAJIMA 1989b).

These circumstances are likely to lead to a "star genealogy" with all sequences separated by an approximately equal amount of evolutionary time. In a star genealogy, $\delta_{ij} = 1$ if and only if the two mutations occur on the same branch. This quantity equals $1/(n-1)^2$ if the two mutations occur on different branches. Weighing these two values by their relative likelihood, the expected value of δ_{ij}, and also of Z_{nS}, is $1/(n-1)$. If $n > 2$, the probability that Z_{nS} equals 1 is $(1/n)^{S-1}$. Despite that each lineage of a star genealogy goes back in time, unbifurcated, directly to the common ancestor of the entire sample, there is no critical region. Thus, we expect smaller values of Z_{nS} under a star genealogy than under a neutral genealogy.

Selection at linked sites will affect the genealogy of a neutral sequence in a different way if recombination occurs between the selected site and the neutral sequence. If recombination is infrequent, we expect that most sequences within a sample will have a recent common ancestor in the sequence that was initially linked to the selectively favored mutation. However, if a recombination event occurs between the selected site and the neutral sequence and this recombination occurred during the sojourn of the beneficial mutation, the most recent common ancestor between this recombinant sequence and the rest of the sample may be far more ancient (Figure 1C; see also Figure 2 in BRAVERMAN *et al.* 1995). Most sequences in the sample of Figure 1C are closely related to each other (linked by a star genealogy). However, the sample also contains one (or a few) sequences that are distantly related to the entire set of sequences in the star genealogy. This will significantly inflate Z_{nS} because all mutations over a high proportion of the genealogy will yield $\delta_{ij} = 1$ (there is large critical region of the genealogy). Thus, while selective sweeps with no recombination (Figure 1B) should reduce Z_{nS}, sweeps with recombination may give Z_{nS} values that exceed the neutral expectation.

Balancing selection of two alternative alleles at a linked locus (Figure 1D) will produce a genealogy that is qualitatively different from the others in Figure 1. Sequences within an "allelic class," where allele refers to the alternatives at the selected locus, are likely to have a recent common ancestor (the top or bottom set of sequences in Figure 1D). However, recombination between the neutral sequence and the selected site is necessary for coalescence of sequences from distinct allelic classes. This is likely to take much longer. Again, because the topology is dominated by the period where only two sequences are present, Z_{nS} is likely to be high.

THEORY

The cumulative distribution function of Z_{nS} can be investigated by conditioning on the genealogy of the sample

$$\text{Prob}[Z_{nS} < c] = \sum_G \text{Prob}[Z_{nS} < c|G]\ \text{Prob}[G], \quad (5)$$

where c ranges from 0 to 1 and G denotes the characteristics of the sample genealogy including branch lengths and topology. The sum is taken over all possible genealogies and Prob[G] denotes the likelihood of any specific genealogy. Prob[$Z_{nS} < c|\ G$] is the probability that $Z_{nS} < c$ given the genealogy G.

This probability is approximated by averaging over a large number of simulations:

$$\text{Prob}[Z_{nS} < c] \approx \frac{1}{W}\sum_{i=1}^{W} Ind[Z_{nS}(i) < c], \quad (6)$$

where $Z_{nS}(i)$ is the calculated value of Z_{nS} from the ith simulation and W is the number of simulations. $Ind[Z_{nS}(i) < c]$ is a function that equals 1 if $Z_{nS}(i) < c$ and equals 0 if not.

Each simulation of evolution was obtained from the following algorithm based on the general methodology described by HUDSON (1990, 1993): (1) establish a random genealogy via simulation (see below), (2) randomly drop S mutations onto the genealogy, (3) determine the mutations present on each sample sequence given the position of mutations on the genealogy, (4) calculate Z_{nS} and store result. A random genealogy was established by first simulating the branch lengths, the T_j in Equation 4, by drawing exponential random numbers with the appropriate parameter. The topology was established by randomly coalescing lineages until the common ancestor of the entire sample was obtained. The topological information was stored in an array that noted, for each branch in the genealogy, whether or not it is ancestral to each of the n sample sequences (HUDSON 1990). Mutations occurring on a branch will be present on each sample sequence to which it is ancestral.

TABLE 1

Upper bounds for z_{nS}

	Sample size (n)																		
S	3	4	5	6	7	8	9	10	12	14	16	18	20	25	30	40	50	60	80
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.90	0.73	0.66	0.58	0.51
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.94
4	1	1	1	1	1	1	1	1	0.86	0.85	0.74	0.73	0.67	0.61	0.58	0.54	0.53	0.52	0.51
	1	1	1	1	1	1	1	1	1	1	1	1	1	0.92	0.83	0.74	0.66	0.62	0.58
5	1	1	1	1	1	1	0.85	0.78	0.72	0.68	0.66	0.65	0.64	0.62	0.61	0.61	0.59	0.55	0.47
	1	1	1	1	1	1	1	1	1	0.90	0.85	0.81	0.78	0.71	0.66	0.64	0.62	0.62	0.61
6	1	1	1	1	0.86	0.81	0.79	0.73	0.71	0.69	0.68	0.68	0.67	0.63	0.57	0.50	0.47	0.45	0.43
	1	1	1	1	1	1	1	0.89	0.83	0.77	0.75	0.72	0.71	0.69	0.68	0.67	0.66	0.62	0.55
7	1	1	1	0.86	0.80	0.78	0.76	0.74	0.73	0.70	0.68	0.63	0.60	0.55	0.52	0.50	0.49	0.48	0.44
	1	1	1	1	1	0.89	0.88	0.85	0.79	0.75	0.74	0.73	0.72	0.70	0.67	0.61	0.57	0.53	0.51
8	1	1	1	0.86	0.80	0.78	0.76	0.76	0.71	0.66	0.62	0.60	0.58	0.55	0.54	0.50	0.45	0.43	0.40
	1	1	1	1	0.90	0.86	0.82	0.81	0.78	0.76	0.76	0.75	0.73	0.65	0.62	0.57	0.55	0.54	0.53
9	1	1	0.88	0.83	0.81	0.79	0.77	0.73	0.67	0.64	0.61	0.60	0.59	0.54	0.51	0.46	0.45	0.43	0.41
	1	1	1	1	0.88	0.84	0.82	0.80	0.79	0.78	0.73	0.67	0.67	0.62	0.60	0.58	0.56	0.51	0.47
10	1	1	0.88	0.83	0.81	0.78	0.73	0.70	0.66	0.64	0.62	0.59	0.57	0.53	0.50	0.47	0.44	0.42	0.38
	1	1	1	0.91	0.86	0.83	0.82	0.81	0.78	0.73	0.69	0.67	0.65	0.63	0.62	0.55	0.51	0.50	0.48
12	1	1	0.87	0.85	0.79	0.74	0.71	0.70	0.66	0.61	0.59	0.57	0.56	0.52	0.49	0.45	0.43	0.41	0.37
	1	1	0.91	0.88	0.86	0.84	0.81	0.77	0.72	0.70	0.69	0.65	0.64	0.60	0.57	0.54	0.50	0.48	0.45
14	1	0.91	0.87	0.81	0.77	0.74	0.72	0.67	0.64	0.61	0.58	0.55	0.54	0.51	0.48	0.44	0.41	0.40	0.36
	1	1	0.92	0.88	0.86	0.81	0.78	0.76	0.73	0.70	0.66	0.64	0.63	0.59	0.55	0.52	0.49	0.47	0.44
16	1	0.92	0.88	0.81	0.77	0.72	0.69	0.67	0.63	0.60	0.57	0.56	0.53	0.49	0.47	0.43	0.40	0.39	0.35
	1	1	0.93	0.88	0.83	0.80	0.77	0.76	0.70	0.68	0.66	0.63	0.61	0.57	0.54	0.50	0.48	0.46	0.42
18	1	0.93	0.86	0.81	0.76	0.72	0.69	0.66	0.62	0.59	0.56	0.54	0.52	0.49	0.46	0.43	0.40	0.38	0.35
	1	1	0.91	0.89	0.83	0.80	0.76	0.73	0.70	0.67	0.64	0.62	0.60	0.56	0.53	0.50	0.47	0.45	0.42
20	1	0.92	0.85	0.81	0.75	0.72	0.68	0.65	0.62	0.58	0.56	0.53	0.51	0.48	0.45	0.42	0.39	0.38	0.35
	1	1	0.91	0.86	0.82	0.79	0.75	0.73	0.69	0.66	0.64	0.60	0.59	0.55	0.52	0.49	0.46	0.44	0.42
25	1	0.93	0.85	0.79	0.74	0.71	0.67	0.65	0.60	0.57	0.54	0.53	0.51	0.47	0.45	0.41	0.38	0.37	0.34
	1	0.95	0.91	0.85	0.81	0.78	0.75	0.72	0.67	0.64	0.62	0.60	0.58	0.54	0.51	0.48	0.45	0.43	0.41
35	1	0.91	0.83	0.78	0.73	0.69	0.66	0.64	0.59	0.56	0.54	0.52	0.50	0.46	0.44	0.40	0.37	0.35	0.33
	1	0.95	0.89	0.84	0.79	0.75	0.72	0.70	0.66	0.63	0.60	0.58	0.56	0.52	0.50	0.46	0.43	0.41	0.38
50	1	0.90	0.83	0.77	0.72	0.68	0.65	0.63	0.58	0.56	0.53	0.51	0.50	0.46	0.44	0.40	0.37	0.35	0.33
	1	0.94	0.88	0.83	0.79	0.75	0.71	0.69	0.65	0.62	0.59	0.57	0.55	0.51	0.49	0.45	0.42	0.40	0.37

The top number in each position is F_{95} (~95% of simulations were below the listed for that n and S) and the bottom number is $F_{97.5}$.

This distribution of Z_{nS} is characterized by F_{95} and $F_{97.5}$, where ~5 and 2.5% of the simulated values of Z_{nS} are above these bounds, respectively (Table 1). It is probably most appropriate to consider Z_{nS} a "one-tailed" test of the neutral model [higher than expected values for Z_{nS} suggest some form of selection at linked sites (Figure 1), whereas lower than expected values of may be caused by nothing more than intragenic recombination]. In a one-tailed test, the 95th and 97.5th percentiles denote P values of 0.05 and 0.025, respectively. Table 1 gives these percentiles for n ranging from 2 to 80 and S ranging from 2 to 50. A library of the full cumulative distribution functions for each case (from which P values can be assigned to any specific value of Z_{nS}) has been compiled on computer files and are available from the author upon request.

The distribution of Z_{nS} is typically skewed with its most likely values less than the mean (Figure 2). The distribution is not smooth and often exhibits "spikes" at specific values, especially when S is small (compare Figure 2, A–C). These small scale changes in probability are not caused by sampling error associated with the limited number of simulated genealogies (they emerge at the same points in distinct sets of simulations with the same values for n and S).

The position of these discontinuities can be pre-

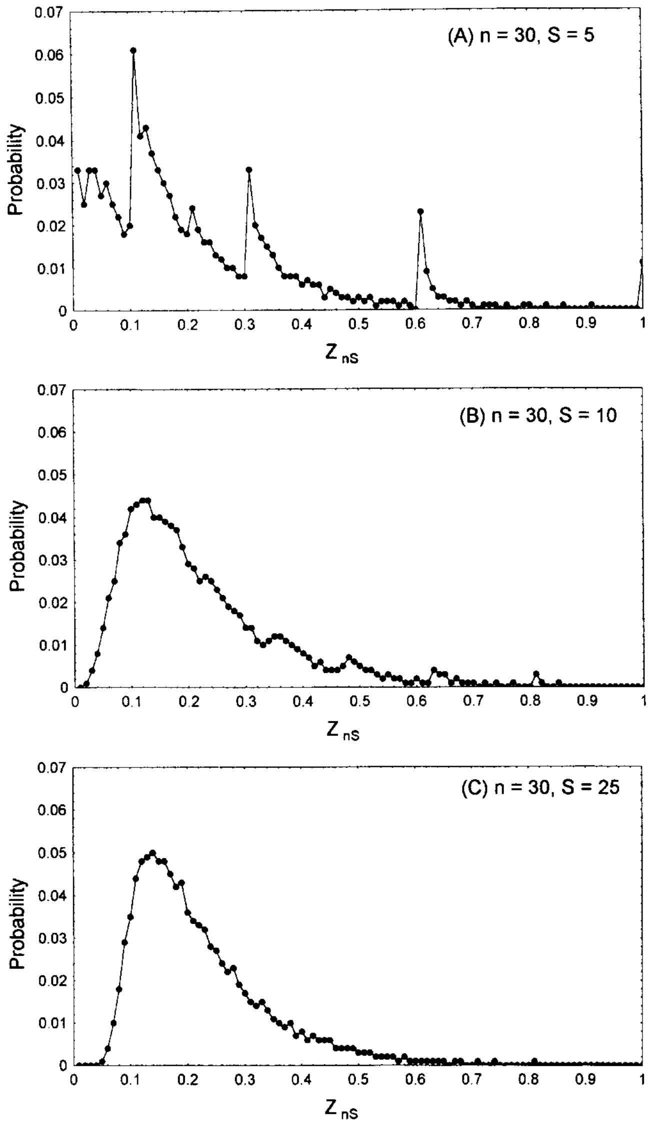

FIGURE 2.—The estimated probability that Z_{nS} equals the value on the x-axis (falls within a 0.01 interval around that point) when $n = 30$ and (A) $S = 5$, (B) $S = 10$, or (C) $S = 25$.

dicted by considering the proportion of pairwise comparisons in which $\delta_{ij} = 1$. All comparisons between mutations that occur in the critical region of genealogy give $\delta_{ij} = 1$. Imagine that all other comparisons between mutations give very small values for δ_{ij}. Under this condition, the distribution of Z_{nS} will have spikes at points corresponding to genealogies with k of the S mutations occurring in the critical region of the genealogy. The value of Z_{nS} given k mutations in the critical region will be slightly larger than $k(k-1)/[S(S-1)]$. For example, with $S = 5$, spikes in the distribution of Z_{nS} are expected to occur at 0.1, 0.3, 0.6, and 1, corresponding to two, three, four, and five mutations in the critical region, respectively.

The spikes predicted by this extreme model correspond closely to those observed for the actual distribution of Z_{nS} in many cases. For example, the distribution with $n = 30$ and $S = 5$ has spikes at each of the points predicted by the extreme model (Figure 2A). With higher numbers of polymorphic sites (*e.g.*, Figure 2, B and C), the spikes corresponding to high numbers of mutations in the critical region are discernable, but spikes at lesser values are smoothed away.

TABLE 2

The percentage of type I errors in 20,000 simulations of evolution for given values of n and θ

	Sample size (n)			
θ	5	10	25	50
$5/f_n$	1.9	4.4	5.1	5.3
$10/f_n$	3.7	5.4	5.2	5.5
$20/f_n$	4.7	5.1	5.3	5.5

Here $f_n = \Sigma 1/i$ for i ranging from 1 to $n - 1$.

Simulation testing of upper bounds: The objective of the prior simulations was to determine the distribution of Z conditional on n and S. The first step of the simulation algorithm was to generate a genealogy by the standard coalescent method given a sample of size n (KINGMAN 1982). The second step, where S mutations are randomly sprinkled onto the sample genealogy, is the means by which the distribution is conditioned on S. This procedure is based on the idea that S alone provides no information about genealogy (HUDSON 1993).

To determine whether F_{95}, as determined in this way, represents a valid test statistic, I conducted a second set of simulations. Specifically, we need to determine if a population of a given size and mutation rate that is undergoing neutral evolution produces data that is inconsistent with the neutral model only 5% of the time. In this set of simulations, the sample size and scaled mutation rate, θ, were specified in advance (not the number of polymorphic sites). Given these two quantities, standard coalescent simulations were performed by the following algorithm: (1) establish a random genealogy via simulation, (2) simulate the number of mutations on the genealogy given θ and the total branch length of the simulated genealogy (T_{cum}), (3) randomly drop mutations onto the genealogy, (4) determine the mutations present on each sample sequence given the position of mutations on the genealogy, (5) calculate Z_{nS} and store with S value. In contrast to the previous procedure, each simulation may differ from others in S as well as Z_{nS}.

I performed 20,000 simulations for each of a range of values of n and θ (Table 2). All simulations where there were less than two polymorphic sites were discarded because Z_{nS} is undefined for these cases (and would not be used on such data). The set of simulations for each value of n and θ were subdivided by S and, for each S, the fraction of "false positives" were calculated (Z_{nS} values above F_{95}). The overall false positive percentage is an average over all values of S weighted by the fraction of simulations resulting in S polymorphic sites.

In most cases, the false positive rate was ~5% (Table 2). However, for the lowest values of n and θ, the percentage was substantially less (the test becomes too conservative). The low rate of false positives is due to the fact that, with low values of n and S, no result is inconsistent with the neutral model (Table 1). For example, with $n = 5$ and $\theta = 5/f_n = 2.40$, 16,447 of 20,000 simulations yielded two to eight polymorphic sites for which even $Z_{nS} = 1$ is not inconsistent with the neutral model.

This second set of simulations generally support F_{95} as an upper bound for testing of the neutral model. This provides some justification for the general method suggested by HUDSON (1993) for generating distributions conditional on S: simulate a random genealogy and then sprinkle on S mutations. I have also investigated two alternative methods for estimating Prob$[Z_{nS} < c]$. In these methods, probabilities were determined by conditioning on θ as well as on S. First, a standard coalescent simulation was performed with a random number of mutations. All simulations that did not result in exactly S mutations were discarded (which is the means by which conditioning affects the distribution). I found that for a given value of S, the expected value of Z_{nS} tends to decrease as θ increases. The first method, described by BERGER and BOOS (1994) and employed in coalescent models by SIMONSEN *et al.* (1995), is to perform the simulations with the lowest value of θ that is consistent with S. A second method, which can be justified from BAYES theorem (J. K. KELLY, unpublished results), involves simulating evolution over a range of values of θ. The distribution of Z_{nS} was then obtained by averaging the conditional probabilities, Prob$[Z_{nS} < c \mid S, \theta = x]$, weighted by the relative likelihood of observing S polymorphic sites given that $\theta = x$. However, I found that neither of these alternative methods perfomed as well the simpler procedure described above. The BERGER and BOOS (1994) method yields F_{95} that are too conservative. The Bayesian model typically yielded false positive rates that exceeded 5%.

APPLICATIONS TO GENETIC DATA FROM *D. melanogaster*

Yellow-ac-scute region: MARTIN-CAMPOS *et al.* (1992) surveyed 10 *D. melanogaster* populations (seven in Europe, two in the USA, and one in Japan) for restriction site variation in a 23.1-kb region on the X chromosome and identified 14 polymorphic sites. This area, the *y-ac-sc* region, is close to the telomere and is known to have a very low rate of recombination (DUBININ *et al.* 1937; BEECH and LEIGH-BROWN 1989). Four of the seven European populations had sufficient variation ($S > 1$) to calculate Z_{nS} and three of four have significantly higher values of Z_{nS} than expected under the neutral model (Table 3). Both American populations have Z_{nS} values that are slightly higher than expected but well within range consistent with the neutral model (Table 3). Finally, the Japanese population has the highest possible value for the test statistic, $Z_{nS} = 1$. However, given the small sample size ($n = 8$) and number of polymorphic sites ($S = 3$), we expect this result under the neutral model ~14% of the time.

TABLE 3

Summary of restriction site data

Population	n	S	Z_{nS}	P
1. Groningen, Holland	25	7	0.33	0.21
2. Canary Islands	25	7	0.70	0.029
3. Barcelona, Spain	50	7	0.52	0.041
4. Huelva, Spain	23	5	1.00	0.015
5. Texas, United States	27	8	0.27	0.28
6. North Carolina, United States	20	8	0.34	0.24
7. Fukuoka, Japan	8	3	1.00	0.14

From MARTIN-CAMPOS *et al.* (1992). P denotes the fraction of simulations equal to or greater than observed value of Z_{nS}. The populations from Requena, Oviedo, and Leon did not have $S > 1$ and are not included here.

***Adh* locus:** A balanced polymorphism of alternative electrophoretic alleles in the Alcohol Dehydrogenase gene (*Adh*) of *D. melanogaster* has been suggested by geographical (OAKESHOTT *et al.* 1982), population genetic (HUDSON *et al.* 1987; KREITMAN and HUDSON 1991), and biochemical analyses (AQUADRO *et al.* 1986; LAURIE *et al.* 1991; LAURIE and STAMS 1994). A collection of *Adh* sequences is presented in Figure 3. Sequences numbered 1–6 and 10–14 were obtained by KREITMAN (1983). Sequences numbered 7–9 and 15 were obtained by LAURIE *et al.* (1991). Finally, the alleles numbered 16–18 are previously unpublished sequences generously provided by MARTIN KREITMAN.

Figure 3 lists all polymorphic sites in this sample within the third intron and the translated region of the fourth exon (which includes the allozyme polymorphism) of the *Adh* gene. The position numbers in Figure 3 follow the scheme in Figure 2 of LAURIE *et al.* (1991) and increase from the 5′ to 3′ direction. The consensus sequence is derived from KREITMAN (1983). The allozyme polymorphism is located at site 1490 where the consensus nucleotide A denotes the "slow" allele and the C nucleotide denotes the "fast" allele.

The *Adh* gene is on the second chromosome of *D. melanogaster* and there is a higher rate of recombination in this area than in the *y-ac-sc* region. Thus, we expect that linkage disequilibrium will decay more rapidly with distance in *Adh*. For this reason, an "expanding window" analysis was applied to the sequence data. The statistic Z_{nS} was calculated by contrasting sequence variation within a window of sites around the fast/slow polymorphism (*e.g.*, HUDSON and KAPLAN 1988; KREITMAN and HUDSON 1991), where a window of width $2k$ in-

		Sequence number																	
Position	C	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1354	g	c	c	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1362	g	a	a	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1387	c	.	.	.	.	.	.	.	g	.	.	.	.	.	.	.	.	.	.
1388	a	.	.	.	.	.	.	.	.	.	.	.	.	.	g	.	.	.	.
1399	g	.	.	.	.	.	.	.	a	.	.	.	.	.	.	.	.	.	.
1400	a	t	t	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1402	t	.	.	.	.	.	.	.	g	.	.	.	.	.	.	.	.	.	.
1405	t	a	a	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1425	c	a	a	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1431	t	c	c	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1443	c	.	.	.	.	.	.	.	.	g	g	g	g	g	g	g	g	.	.
1452	c	.	.	.	.	.	t	t	.	t	t	t	t	t	t	t	t	.	.
1490	a	.	.	.	.	.	.	.	.	c	c	c	c	c	c	c	c	.	.
1518	c	.	.	.	.	.	t	t	.	t	t	t	t	t	t	t	t	.	.
1527	t	.	.	.	.	.	c	c	.	c	c	c	c	c	c	c	c	.	.
1555	c	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	t	.	.
1557	a	.	.	.	.	.	c	c	.	c	c	c	c	c	c	c	c	.	.
1596	g	.	.	a	a	.	a	.	a	.	.	.	.	.	.	.	.	a	a
1630	t	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	c

FIGURE 3.—Sequence variation near the fast/slow polymorphism in the *Adh* gene of *D. melanogaster*. The C column denotes the concensus sequence. Sites equivalent to the concensus are denoted by (.) and differences are denoted by the substituted base.

cludes all polymorphisms within k sites of the fast/slow polymorphism. The width k was progressively expanded from 0 to 140 sites and the statistics recalculated whenever a new polymorphism was included.

Each point in Figure 4 denotes the value of Z_{nS} in a window of the width given by the x-axis. High values of Z_{nS} are observed for small windows around the fast/slow polymorphism, but Z_{nS} decays rapidly with window size. The windows, including four and five polymorphisms, respectively ($Z_{n4} = 0.82$, $Z_{n5} = 0.78$; denoted by two asterisks in Figure 4), are significantly higher than expected (the fraction of simulations equal to or greater than the observed values were 0.04 and 0.03, respectively). The windows including three or six polymorphisms ($Z_{n3} = 0.76$, $Z_{n6} = 0.57$; denoted by one asterisk in Figure 4) are borderline significant ($P = 0.08$ for each).

DISCUSSION

There is considerable interest in the patterns of genetic variability within genomic regions of low recombination (AGUADE *et al.* 1989; STEPHAN and LANGLEY 1989; BEGUN and AQUADRO 1992; CHARLESWORTH *et al.* 1993; CHARLESWORTH 1994). The distribution of Z_{nS} obtained here is based on the assumption of no recombination between sites. This assumption is also essential to other single population/single locus tests of neutrality (WATTERSON 1978; TAJIMA 1989a; FU and LI 1993; BRAVERMAN *et al.* 1995; SIMONSEN *et al.* 1995). The expected evolutionary patterns under the extreme condition of no recombination provide a basis for comparison when more complicated and realistic models that include recombination are considered.

In the absence of recombination, Z_{nS} is a measure of allele frequency equivalency across polymorphic sites. This measure declines in value as asymmetry among loci increases. The neutral model predicts a certain level of allele frequency asymmetry among polymorphic sites. However, when natural selection acts on a polymorphism that is closely linked to neutral sites, allele frequency asymmetries may be reduced. For this reason, higher than expected values of Z_{nS} may represent a molecular signature of natural selection.

Unfortunately, relatively high values of Z_{nS} are necessary to reject the neutral model (Table 1). The high variance of Z_{nS} is a consequence of the stochastic nature of sample genealogies under neutrality. This also limits the power of other single population/single locus tests. While such tests may not be very powerful alone, their combined application may prove quite useful, especially if different tests are sensitive to distinct types of deviation from the neutral model (see next section). Direct studies of specific evolutionary models (*e.g.*, BRAVERMAN *et al.* 1995; SIMONSEN *et al.* 1995) are required to assess the statistical power of Z_{nS} relative to other tests (*e.g.*, TAJIMA 1989a; FU and LI 1993) in different biological circumstances.

Applications: There have been numerous surveys of restriction site variation in the *yellow-ac-scute* region of *D. melanogaster* (AGUADE *et al.* 1989; BEECH and LEIGH-

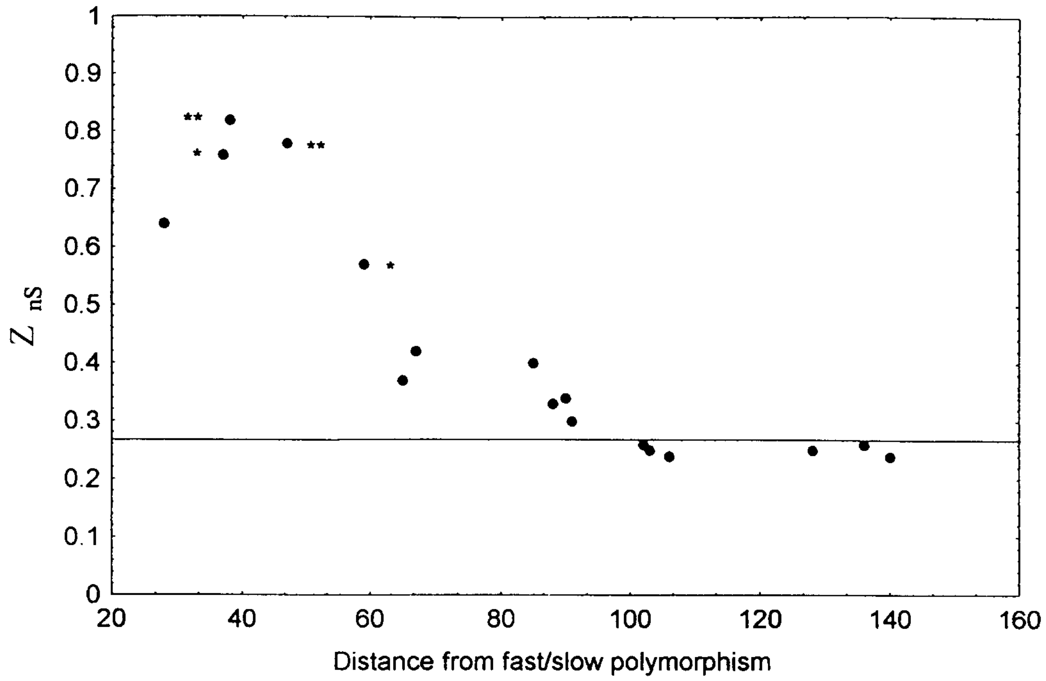

FIGURE 4.—The value of Z_{nS} within a window of sites around the fast/slow polymorphism of *Adh* as a function of the distance between site 1490 and the edge of the window. Each point represents a window width at which a new polymorphic site is included. Significantly higher than expected values of Z_{nS}: **0.01 < P < 0.05 and *0.05 < P < 0.10.

BROWN 1989; EANES *et al.* 1989; MACPHERSON *et al.* 1990; BEGUN and AQUADRO 1991; MARTIN-CAMPOS *et al.* 1992). Three general features of these data are (1) reduced levels of nucleotide variation relative to other genomic regions in *D. melanogaster*, (2) an excess of rare alleles at polymorphic sites, and (3) linkage disequilibrium among polymorphic sites.

BEGUN and AQUADRO (1991) and MARTIN-CAMPOS *et al.* (1992) suggest that selectively advantageous mutations have recently occurred in the *y-ac-sc* region with hitch-hiking (*sensu* MAYNARD SMITH and HAIGH 1974) affecting the patterns of genetic variation at linked sites. This contention is supported by two aspects of the MARTIN-CAMPOS *et al.* (1992) data. The first is based on the observed allele frequencies at polymorphic sites. In the four European populations with more than one polymorphic site, there is an excess of rare alleles relative to the neutral expectation (Tajima's D is negative in each case, with statistically significant values for the Barcelona and Huelva populations).

Second, hitch-hiking reduces the level of intraspecific neutral variation by reducing T_{cum}, the total length of the sample genealogy. It should not affect the expected divergence among species however (KIMURA 1983; HUDSON *et al.* 1987). The level of nucleotide variation within the *yellow-ac* region is lower than expected given the level of interspecific divergence between *D. melanogaster* and either *D. simulans* or *D. sechellia* under neutrality (BEGUN and AQUADRO 1991; MARTIN-CAMPOS *et al.* 1992). This observation supports a hitch-hiking model.

The present study extends these analyses by considering whether or not the associations between polymorphic sites observed in the data of MARTIN-CAMPOS *et al.* (1992) can be explained by mutation-drift balance among linked sites. Higher than expected values for Z_{nS} were observed in the European samples from Barcelona, Huelva, and the Canary islands (Table 3). Thus, the pattern of associations among polymorphic loci is also inconsistent with a neutral model.

These calculations of Z_{nS} complement the previous application of Tajima's test to these data [table 7 in MARTIN-CAMPOS *et al.* (1992)]. For example, the rarer allele at each of the five polymorphic sites in the Huelva sample appears in only one of the 23 sequences. This yields the most negative value possible for Tajima's D (given five polymorphic sites) and suggests a sample genealogy like Figure 1, B or C. However, it is notable that all of the sequence variation in the Huelva sample is concentrated on a single sample allele [haplotype 38 in Table 4 of MARTIN-CAMPOS *et al.* (1992)]. This pattern of interlocus association among polymorphisms is indicated by the significantly high value of Z_{nS} for this sample and suggests that the genealogy in Figure 1B is much less likely than the alternative model allowing recombination between selected and neutral loci (Figure 1C).

Selective neutrality has also been rejected in previous analyses of the *Adh* gene of *D. melanogaster* (HUDSON *et al.* 1987; KREITMAN and HUDSON 1991). These analyses have contrasted the amount of polymorphism within *D. melanogaster* with the amount of divergence between *D. melanogaster* and closely related species in the *Adh* region. The amount of polymorphism around site 1490 is greater than expected, which suggests a balanced polymorphism.

Linkage disequilibrium is expected among selectively neutral polymorphic sites closely linked to a balanced polymorphism and strong linkage disequilibrium is observed close to the fast/slow polymorphism in *Adh* (KREITMAN 1983). However, tests demonstrating significant linkage disequilibrium between sites generally do not indicate the cause of interlocus associations (LEWONTIN 1995). Linkage disequilibrium is expected among closely linked sites as a result of mutation/drift balance without any contribution from selection (OHTA and KIMURA 1969, 1971; WEIR and COCKERHAM 1974; GRIFFITHS 1981).

The substantial pattern of allele frequency equivalency across polymorphic sites near the fast/slow site in *Adh* is perhaps more informative. For example, the mutant allele has the same frequency at the sites numbered 9–13 and 15 in Figure 5. This equivalency across sites yields values of Z_{nS} within small windows around site 1490 that are too large to be consistent with neutrality (Figure 6).

The rapid decay of Z_{nS} values with increasing window size in Figure 6 indicates the sensitivity of this measure to recombination. This will limit the utility of Z_{nS} as a test unless there is a very low rate of recombination (as in mitochondrial DNA or telomeric regions of the nuclear genome) or a large number of polymorphic sites in close proximity to each other (as in *Adh*).

With intragenic recombination, Z_{nS} will be determined only partially by the sample genealogy. For this reason, it may prove useful to modify the statistic to hedge against the effects of recombination. One potential method involves contrasting Z_{nS}, which is sensitive to both recombination and genealogy, with another statistic that only depends on recombination. A standardized measure of linkage disequilibrium that is relatively independent of allele frequency was proposed by LEWONTIN (1964):

$$D'_{ij} = \frac{D_{ij}}{\mathrm{Min}\,[p_i(1-p_j),\, p_j(1-p_i)]}, \quad \text{if } D_{ij} > 0$$

$$D'_{ij} = \frac{D_{ij}}{\mathrm{Min}\,[p_i p_j,\, (1-p_i)(1-p_j)]}, \quad \text{if } D_{ij} < 0, \qquad (7)$$

where Min [a,b] denotes the smaller of a and b. Lewontin's D' ranges from -1 to 1 and assumes intermediate values only when all four two-locus haplotypes are present in a sample {00, 01, 10, and 11}. When the two polymorphic sites are completely linked and derived from a single ancestor, there can be at most three of these haplotypes present. Thus, in the absense of recombination, $(D'_{ij})^2$ equals 1 for all comparisons between mutations and is insensitive to genealogy.

Two potential measures of interlocus association that compare the squared correlation among sites and $(D'_{ij})^2$ are as follows:

$$Z^*_{nS} = Z_{nS} + 1 - D^*_{nS} \qquad (8)$$

and

$$Z^{**}_{nS} = \frac{Z_{nS}}{D^*_{nS}}, \qquad (9)$$

where D^*_{nS} is the averaged squared value of Lewontin's D':

$$D^*_{nS} = \frac{2}{S(S-1)} \sum_{i=1}^{S-1} \sum_{j=i+1}^{S} (D'_{ij})^2. \qquad (10)$$

The distributions of Z^*_{nS} and Z^{**}_{nS} are equal to the distribution of Z_{nS} when sites within the neutral sequence are completely linked because then D^*_{nS} must equal 1. When intragenic recombination occurs, we expect that Z^*_{nS} and Z^{**}_{nS} will be greater than Z_{nS}. These modified statistics will be useful if the genealogical information provided by Z_{nS} when there is no intragenic recombination is preserved in Z^*_{nS} and Z^{**}_{nS} when intragenic recombination does occur.

Patterns of linkage disequilibria are now being used to isolate disease loci in humans (FEDER *et al.* 1996; LITTLE 1996). Present analyses are largely nonstatistical and do not distinguish between the possible causes for linkage disequilibria. The "sliding window" methodology developed by HUDSON and KAPLAN (1988) provides a statistical means to isolate selected loci based on the amount of variation within a genomic region. The present study suggests that a similar method based on the patterns of interlocus associations within a region may also prove a useful tool of genetic analysis.

This article benefitted from comments by C. BERGMAN, M. WADE, D. KELLY, K. F. MALLOY, R. HUDSON, and an anonymous reviewer, and from discussions with J. THORNE, M. KREITMAN, E. STALH, B. CHARLESWORTH, D. CHARLESWORTH, and M. NORDBORG. The author received financial support from a National Institutes of Health (NIH) genetics training grant, a NIH National Research Service Award postdoctoral grant, and a Hughes Foundation postdoctoral award.

LITERATURE CITED

AGUADE, M., N. MIYASHITA and C. H. LANGLEY, 1989 Reduced variation in the *yellow-achaete-scute* region in natural populations of *Drosophila melanogaster.* Genetics **122:** 607–615.

AUADRO, C. F., S. F. DESSE, M. BLAND, C. H. LANGLEY and C. C. LAURIE-AHLBERG, 1986 Molecular population genetics of the *alcohol dehydrogenase* gene region of *Drosophila melanogaster.* Genetics **114:** 1165–1190.

BEECH, R. N., and A. J. LEIGH-BROWN, 1989 Insertion-deletion variation at the *yellow, achaete-scute region* in two natural populations of *Drosophila melanogaster.* Genet. Res. **53:** 7–15.

BEGUN, D. J., and C. F. AQUADRO, 1991 Molecular population genetics of the distal portion of the X chromosome in Drosophila: evidence for genetic hitchhiking of the *yellow-achaete region.* Genetics **129:** 1147–1158.

BEGUN, D. J., and C. F. AQUADRO, 1992 Levels of naturally occuring DNA polymorphism correlate with recombination rates in *D. melanogaster.* Nature **356:** 519–520.

BERGER, R. L., and D. D. BOOS, 1994 P values maximized over a confidence set for the nuisance parameter. J. Am. Stat. Assoc. **89:** 1012–1016.

BRAVERMAN, J. M., R. R. HUDSON, N. L. KAPLAN, C. H. LANGLEY and W. STEPHAN, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140:** 783–796.

CHARLESWORTH, B., 1994 The effect of background selection against deleterious mutations on weakly-selected, linked variants. Genet. Res. **63:** 213–227.

CHARLESWORTH, B., M. T. MORGAN and D. CHARLESWORTH, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289–1303.

DUBININ, N. P., N. N. SOKOLOV and G. G. TINIAKOV, 1937 Crossover between the genes "*yellow,*" "*achaete,*" and "*scute.*" Dros. Inf. Serv. **8:** 76.

Eanes, W. F., J. Labate and J. W. Ajioka, 1989 Restriction-map variation with the *yellow-achaete-scute* region in five populations of *Drosophila melanogaster.* Mol. Biol. Evol. **6:** 492–502.

Ewens, W. J., 1990 Population genetics theory—the past and the future, pp. 177–227 in *Mathematical and Statistical Developments of Evolutionary Theory,* edited by S. Lessard. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Feder, J. N., A. Gnirke, W. Thomas, Z. Tsuchihashi *et al.,* 1996 A novel *MHC class I*-like gene is mutated in patients with hereditary haemochromatosis. Nat. Genet. **13:** 399–408.

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

Griffiths, R. C., 1981 Neutral two-locus multiple allele models with recombination. Theor. Popul. Biol. **19:** 169–186.

Hartl, D. L., and A. G. Clark, 1989 *Principles of Population Genetics.* Sinauer Associates, Sunderland, MA.

Hudson, R. R., 1990 Gene geneologies and the coalescent process. Oxford Surv. Evol. Biol. **7:** 1–44.

Hudson, R. R., 1993 The how and why of generating gene genealogies, pp. 23–36 in *Mechanisms of Molecular Evolution,* edited by N. Takahata and A. G. Clark. Sinauer Associates, Sunderland, MA.

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. Aguade, 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.

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The hitchiking effect revisited. Genetics **123:** 887–899.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.

Kingman, J. F. C., 1982 The coalescent. Stochast. Proc. Appl. **13:** 235–248.

Kreitman, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster.* Nature **304:** 412–417.

Kreitman, M., 1990 Detecting selection at the level of DNA, pp. 204–221 in *Evolution at the Molecular Level,* edited by R. K. Selander, A. G. Clark and T. S. Whittam. Sinauer Associates, Sunderland, MA.

Kreitman, M., and R. R. Hudson, 1991 Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics **127:** 565–582.

Laurie, C. C., and L. F. Stam, 1994 The effect of an intronic polymorphism on alcohol dehydrogenase expression in *Drosophila melanogaster.* Genetics **138:** 379–385.

Laurie, C. C., J. T. Bridgham and M. Choudhary, 1991 Associations between DNA sequence variation and variation in expression of the Adh gene in natural populations of *Drosophila melanogaster.* Genetics **129:** 489–499.

Lewontin, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics **49:** 49–67.

Lewontin, R. C., 1995 The detection of linkage disequilibrium in molecular sequence data. Genetics **140:** 377.

Little, P., 1996 Woman's meat, a man's poison. Nature **382:** 494–495.

Macpherson, J. N., B. S. Weir and A. J. Leigh-Brown, 1990 Extensive linkage disequilibrium in the *achaete-scute* complex of *Drosophila melanogaster.* Genetics **126:** 121–129.

Martin-Campos, J. M., J. M. Comeron, N. Miyashita and M. Aguade, 1992 Intraspecific and interspecific variation in the *y-ac-sc* region of *Drosophila simulans* and *Drosophila melanogaster.* Genetics **130:** 805–816.

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

Oakeshott, J. G., J. B. Gibson, P. R. Anderson, W. R. Knibb, D. G. Anderson *et al.,* 1982 Alcohol dehydrogenase and glycerol-3-phosphate dehydrogenase clines in *Drosophila melanogaster* on different continents. Evolution **36:** 86–96.

Ohta, T., and M. Kimura, 1969 Linkage disequilibrium due to random genetic drift. Genet. Res. **13:** 47–55.

Ohta, T., and M. Kimura, 1971 Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. Genetics **68:** 571–580.

Simonsen, K. L., G. A. Churchill and C. F. Aquadro, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413–429.

Slatkin, M., 1989 Detecting small amounts of gene flow from phylogenies of alleles. Genetics **121:** 609–612.

Stephan, W., and C. H. Langley, 1989 Molecular variation in the centromeric region of the X chromosome in three *Drosophila ananassae* populations. I. Contrasts between the vermilion and forked loci. Genetics **121:** 89–99.

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

Tajima, F., 1989b 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. **52:** 213–222.

Tavare, S., 1984 Line-of-descent and genealogical processes and their applications in population genetics models. Theor. Popul. Biol. **26:** 119–164.

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

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

Weir, B. S. and C. C. Cockerham, 1974 Behavior of pairs of loci in finite monoecious populations. Theor. Popul. Biol. **6:** 323–354.

Communicating editor: R. R. Hudson