Detecting Linkage Disequilibrium in Bacterial Populations

Bernhard Haubold,* Michael Travisano,* Paul B. Rainey* and Richard R. Hudson†

**Department of Plant Sciences, University of Oxford, Oxford OX1 3RB, United Kingdom and †Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92717*

Manuscript received April 2, 1998
Accepted for publication August 21, 1998

ABSTRACT

The distribution of the number of pairwise differences calculated from comparisons between n haploid genomes has frequently been used as a starting point for testing the hypothesis of linkage equilibrium. For this purpose the variance of the pairwise differences, V_D, is used as a test statistic to evaluate the null hypothesis that all loci are in linkage equilibrium. The problem is to determine the critical value of the distribution of V_D. This critical value can be estimated either by Monte Carlo simulation or by assuming that V_D is distributed normally and calculating a one-tailed 95% critical value for V_D, L, $L = E(V_D) + 1.645\sqrt{\text{Var}(V_D)}$, where $E(V_D)$ is the expectation of V_D, and $\text{Var}(V_D)$ is the variance of V_D. If V_D (observed) $> L$, the null hypothesis of linkage equilibrium is rejected. Using Monte Carlo simulation we show that the formula currently available for $\text{Var}(V_D)$ is incorrect, especially for genetically highly diverse data. This has implications for hypothesis testing in bacterial populations, which are often genetically highly diverse. For this reason we derive a new, exact formula for $\text{Var}(V_D)$. The distribution of V_D is examined and shown to approach normality as the sample size increases. This makes the new formula a useful tool in the investigation of large data sets, where testing for linkage using Monte Carlo simulation can be very time consuming. Application of the new formula, in conjunction with Monte Carlo simulation, to populations of *Bradyrhizobium japonicum*, *Rhizobium leguminosarum*, and *Bacillus subtilis* reveals linkage disequilibrium where linkage equilibrium has previously been reported.

BACTERIA might be called "facultative sexuals" because they can exchange genetic material through conjugation, transformation, and transduction, but genetic exchange is not a part of their reproductive mode. Just how frequently recombination takes place in bacteria has been a topic of debate since the first major study of bacterial population genetics, in which *Escherichia coli* genomes were assumed to recombine frequently leading to linkage equilibrium (Milkman 1973). Selander and Levin (1980) showed that this assumption was incorrect and that *E. coli* populations consisted of many asexual clones evolving in genetic isolation from all other clones comprising the species (*cf.* Maruyama and Kimura 1980, but see Guttman and Dykhuizen 1994). During the 1980s this clonal model was thought to hold for all bacterial populations until Istock *et al.* (1992) reported that a local population of *Bacillus subtilis* was in linkage equilibrium and argued that this resulted from frequent mixis. In addition to *B. subtilis*, linkage equilibrium has been reported for *Neisseria gonorrhoeae* (O'Rourke and Stevens 1993), subpopulations of Rhizobium (Souza *et al.* 1992; Maynard Smith *et al.* 1993; Bottomley *et al.* 1994; Strain *et al.* 1995), *Burkholderia cepacia* (Wise *et al.* 1995), *Helicobacter pylori* (Go *et al.* 1996), and fluorescent Pseudomonas (Haubold and Rainey 1996).

The conclusion of linkage equilibrium reached in these studies is based on the variance of the distribution of the number of pairwise differences (V_D) among bacterial isolates that have been subjected to genetic analysis at multiple loci. V_D can be compared to a critical value obtained under the null hypothesis that all loci are in linkage equilibrium. This approach was first developed by Brown *et al.* (1980), who applied it to allozyme data from wild barley, *Hordeum spontaneum*. Whittam *et al.* (1983) pioneered its use in bacterial population genetics, and more recently this method served as the basis for an extensive comparative study of bacterial population structure (Maynard Smith *et al.* 1993).

There are two methods of calculating a critical value for V_D. (1) The null distribution of V_D can be simulated on a computer, and (2) assuming the null distribution of V_D is normal, a critical value can be calculated by the well-known method of adding x standard deviations to $E(V_D)$. But, as it is not known whether the null distribution of V_D is normal, Monte Carlo simulation has recently emerged as the preferred way for testing linkage equilibrium in bacterial populations (Souza *et al.* 1992; Wise *et al.* 1995; Haubold and Rainey 1996). However, this approach is computationally intensive and many workers have preferred to use the simplifying assump-

Corresponding author: Bernhard Haubold, Max-Planck-Institut für Chemische Ökologie, Tatzendpromenade 1a, D-07745 Jena, Germany. E-mail: haubold@ice.mpg.de

tion of normality for hypothesis testing. In this case the correct test depends above all on an accurate estimator of the variance of V_D.

THE TRADITIONAL METHOD OF COMPUTING THE VARIANCE OF V_D

Suppose we have n sampled haploid individuals, arbitrarily numbered from 1 to n, that have been genetically assayed at q loci. Let d_{ij} denote the number of loci at which individuals i and j differ. Then the variance of pairwise differences is by definition equal to

$$V_D = \sum_{i=1}^{n-1}\sum_{j>i}^{n}(d_{ij} - \bar{d})^2 \Big/ \binom{n}{2}, \tag{1}$$

where

$$\bar{d} = \sum_{i=1}^{n-1}\sum_{j>i}^{n} d_{ij} \Big/ \binom{n}{2}. \tag{2}$$

The distribution of V_D depends on how replicate samples would be generated. In this article, we assume that replicate samples are generated by randomly shuffling the original alleles among the sampled haplotypes. In this way, the numbers of alleles and the frequencies of the alleles at individual loci are exactly the same in each replicate as in the original sample, but there is no statistical association of alleles on haplotypes except that which arises by chance. This shuffling method is the method suggested by Souza *et al.* (1992). The distribution of V_D under this randomization is taken to be our null distribution. We note that the distribution of V_D would be slightly different if sampling were done with replacement. Under our randomization scheme the expectation of V_D is

$$E(V_D) = \sum_{j=1}^{r} h_j(1 - h_j), \tag{3}$$

where

$$h_j = \left(\frac{n}{n-1}\right)\left(1 - \sum_i p_{ij}^2\right) \tag{4}$$

and where p_{ij} is the frequency in the sample of the ith allele at the jth locus. We note that h_j is an unbiased estimator of the population genetic diversity.

Brown *et al.* (1980) suggested that the one-tailed 95% critical value for V_D could be calculated assuming that the distribution of V_D is normal. Thus they estimated this critical value by

$$L_{\text{old}} = E(V_D) + 1.645\sqrt{\text{Var}(V_D)_{\text{old}}}, \tag{5}$$

where $\text{Var}(V_D)_{\text{old}}$ is an estimate of the variance of V_D calculated as

$$\text{Var}(V_D)_{\text{old}} = \frac{(n-1)^2}{n^3} m_4 - \frac{(n-1)(n-3)}{n^3} E(V_D)^2 \approx \frac{m_4 - E(V_D)^2}{n}, \tag{6}$$

where

$$m_4 = \sum h_j - 7\sum h_j^2 + 12\sum h_j^3 - 6\sum h_j^4 + 3(\sum h_j - \sum h_j^2)^2 \tag{7}$$

(Brown *et al.* 1980).

In the next section we derive a formula for the variance of V_D under the randomization scheme of Souza *et al.* (1992) and show that (6) is inappropriate for calculating the variance of V_D under these circumstances.

COMPUTING THE VARIANCE OF V_D

In this section we obtain an exact expression for the variance of V_D under the shuffling of alleles across individuals (the sampling without replacement method; see also Hudson 1994). In the following, d_{ij} denotes the random number of loci at which individual i and j differ in a shuffled sample. First we write V_D in terms of s_{ij}, the number of loci at which individuals i and j are identical. Noting that $s_{ij} = q - d_{ij}$, it follows that

$$V_D = \sum_{i=1}^{n-1}\sum_{j>i}^{n}(s_{ij} - \bar{s})^2 \Big/ \binom{n}{2} = \left(\sum_{i=1}^{n-1}\sum_{j>i}^{n} s_{ij}^2 \Big/ \binom{n}{2}\right) - \bar{s}^2, \tag{8}$$

where

$$\bar{s} = \sum_{i=1}^{n-1}\sum_{j>i}^{n} s_{ij} \Big/ \binom{n}{2} = q - \bar{d}. \tag{9}$$

Because under the randomization scheme that we are considering $\bar{s}$ is a constant, it follows that

$$\begin{aligned}
\text{Var}(V_D) &= \text{Var}\left(\sum_{i=1}^{n-1}\sum_{j>i}^{n} s_{ij}^2 \Big/ \binom{n}{2}\right) \\
&= \frac{1}{\binom{n}{2}^2}\sum_{i=1}^{n-1}\sum_{j>i}^{n}\left\{\sum_{k=1}^{n-1}\sum_{l>k}^{n}\text{Cov}(s_{ij}^2, s_{kl}^2)\right\} \\
&= \frac{1}{\binom{n}{2}}\sum_{k=1}^{n-1}\sum_{l>k}^{n}\text{Cov}(s_{12}^2, s_{kl}^2) \\
&= \frac{1}{\binom{n}{2}}\left[\text{Var}(s_{12}^2) + \frac{(n-2)(n-3)}{2}\text{Cov}(s_{12}^2, s_{34}^2) + 2(n-2)\text{Cov}(s_{12}^2, s_{13}^2)\right]
\end{aligned}$$

$$= \frac{1}{\binom{n}{2}}\left[E(s_{12}^4) + \frac{(n-2)(n-3)}{2}E(s_{12}^2 s_{34}^2) + 2(n-2)E(s_{12}^2 s_{13}^2)\right] - E(s_{12}^2)^2, \quad (10)$$

where E denotes expectation under the randomization scheme.

We now proceed to derive expressions for each of the terms on the right-hand side of the last line of Equation 10. Let x_k be an indicator variable, equal to one if individual 1 and individual 2 are identical at locus k, and zero otherwise. Then

$$s_{12} = \sum_{k=1}^{q} x_k \quad (11)$$

and

$$E(s_{12}) = \sum_{k=1}^{q} \phi_k = \bar{s}, \quad (12)$$

where ϕ_k is the probability that two randomly chosen individuals are identical at locus k. For our case,

$$\phi_k = \sum_m p_{mk}(np_{mk} - 1)/(n-1), \quad (13)$$

where p_{mk} is the frequency of the mth allele at the kth locus in the original sample, and the sum is over all alleles at locus k. Similarly,

$$E(s_{12}^2) = E\left[\left(\sum_{k=1}^{q} x_k\right)^2\right] = E\left[\sum_{k=1}^{q} x_k^2 + \sum_i \sum_{j\neq i} x_i x_j\right] = \sum_{k=1}^{q} \phi_k + \sum_i \sum_{j\neq i} \phi_i \phi_j = \bar{s} + \bar{s}^2 - \sum_{i=1}^{q} \phi_i^2. \quad (14)$$

To calculate $E(s_{ij}^4)$, we write

$$E(s_{12}^4) = E\left[\left(\sum_{k=1}^{q} x_k\right)^4\right] = E\left[\sum_{k=1}^{q} x_k^4 + \binom{4}{3}\sum_i \sum_{j\neq i}^{q} x_i^3 x_j + \left[\binom{4}{2}\sum_{i=1}^{q}\sum_{j\neq i}^{q} x_i^2 x_j^2 \Big/ 2\right] + \binom{4}{2}\sum_{i=1}^{q}\sum_{j\neq i}^{q}\sum_{k\neq i,j}^{q} x_i^2 x_j x_k + \sum_{i=1}^{q}\sum_{j\neq i}^{q}\sum_{k\neq i,j}^{q}\sum_{l\neq i,j,k}^{q} x_i x_j x_k x_l\right]$$
$$= \sum_{k=1}^{q} \phi_k + 7\sum_i \sum_{j\neq i}^{q} \phi_i \phi_j + 6\sum_{i=1}^{q}\sum_{j\neq i}^{q}\sum_{k\neq i,j}^{q} \phi_i \phi_j \phi_k + \sum_{i=1}^{q}\sum_{j\neq i}^{q}\sum_{k\neq i,j}^{q}\sum_{l\neq i,j,k}^{q} \phi_i \phi_j \phi_k \phi_i. \quad (15)$$

To arrive at the last line, we have used the fact that an indicator variable to any power is equal to the indicator variable itself. (For example, $x_k^4 = x_k$.) We have also made use of the fact that x_k is independent of x_j, for $j \neq k$. We show later that the double, triple, and quadruple sums on the last line of (15) can be written as single sums and products of single sums of terms involving powers of the ϕ_i's.

Similarly, to calculate the other terms in (10) we define z_k to be one if individuals 3 and 4 are identical at locus k and zero otherwise, and we define y_k to be one if individuals 1 and 3 are identical at locus k. It follows that

$$E(s_{12}^2 s_{34}^2) = E\left[\left(\sum_{k=1}^{q} x_k\right)^2\left(\sum_{k=1}^{q} z_k\right)^2\right] = \sum_i \sum_{j\neq i}^{q} \phi_i \phi_j + 2\sum_{i=1}^{q}\sum_{j\neq i}^{q}\sum_{k\neq i,j}^{q} \phi_i \phi_j \phi_k + \sum_{i=1}^{q}\sum_{j\neq i}^{q}\sum_{k\neq i,j}^{q}\sum_{l\neq i,j,k}^{q} \phi_i \phi_j \phi_k \phi_l + \sum_{i=1}^{q} \Delta_i + 4\sum_{i=1}^{q}\sum_{j\neq i}^{q} \Delta_i \phi_j + 2\sum_{i=1}^{q}\sum_{j\neq i}^{q} \Delta_i \Delta_j + 4\sum_{i=1}^{q}\sum_{j\neq i}^{q}\sum_{k\neq i,j}^{q} \Delta_i \phi_j \phi_k, \quad (16)$$

where Δ_k is the probability that individuals 1 and 2 are identical at locus k and individuals 3 and 4 are also identical at this locus. Recall that alleles are assigned to individuals randomly without replacement, so

$$\Delta_k = \sum_m \left(p_{mk}\frac{(np_{mk}-1)}{n-1}\right)\left(\frac{(np_{mk}-2)}{n-2}\frac{(np_{mk}-3)}{n-3} + \sum_{j\neq m}\frac{np_{jk}}{n-2}\frac{(np_{jk}-1)}{n-3}\right).$$

Similarly,

$$E(s_{12}^2 s_{13}^2) = E\left[\left(\sum_{k=1}^{q} x_k\right)^2\left(\sum_{k=1}^{q} y_k\right)^2\right] = \sum_i \sum_{j\neq i}^{q} \phi_i \phi_j + 2\sum_{i=1}^{q}\sum_{j\neq i}^{q}\sum_{k\neq i,j}^{q} \phi_i \phi_j \phi_k + \sum_{i=1}^{q}\sum_{j\neq i}^{q}\sum_{k\neq i,j}^{q}\sum_{l\neq i,j,k}^{q} \phi_i \phi_j \phi_k \phi_l + \sum_{i=1}^{q} \Gamma_i + 4\sum_{i=1 j\neq i}^{q}\sum^{q} \Gamma_i \phi_j + 2\sum_{i=1 j\neq i}^{q}\sum^{q} \Gamma_i \Gamma_j + 4\sum_{i=1}^{q}\sum_{j\neq i}^{q}\sum_{k\neq i,j}^{q} \Gamma_i \phi_j \phi_k, \quad (17)$$

where Γ_k is the probability that individuals 1, 2, and 3 are identical to each other at locus k,

$$\Gamma_k = \sum_m p_{mk}\frac{(np_{mk}-1)}{n-1}\frac{(np_{mk}-2)}{n-2}.$$

One can now calculate $\mathrm{Var}(V_D)$ using (10) together with (15), (16), and (17).

We can write the results in a way that does not require double, triple, or quadruple sums. For example, note that

$$\sum_i \sum_{j\neq i}^{q} \phi_i \phi_j = \sum_i^{q} \phi_i \sum_{j\neq i}^{q} \phi_j = \sum_i^{q} \phi_i(\bar{s} - \phi_i)$$

$$= \bar{s}\sum_i^q \phi_i - \sum_i^q \phi_i^2 = \bar{s}^2 - \sum_i^q \phi_i^2.$$

In a similar fashion, the other multiple sums can be reduced to terms involving the following single sums:

$$s_k = \sum_i^q \phi_i^k, \quad k = 1, \ldots, 4$$

$$d_k = \sum_i^q \Delta_i^k, \quad k = 1, 2$$

$$g_k = \sum_i^q \Gamma_i^k, \quad k = 1, 2$$

$$D_k = \sum_i^q \Delta_i \phi_i^k, \quad k = 1, 2$$

$$G_k = \sum_i^q \Gamma_i \phi_i^k, \quad k = 1, 2.$$

After some manipulation, the result is

$$\begin{aligned}\mathrm{Var}(V_D) = {} & [4s_3 - s_2 - 6s_4 + 8s_1s_3 - 4s_1s_2 \\ & + 2s_2^2 - 4s_1^2s_2 - 4D_1 + 8D_2 + d_1 + 2d_1^2 \\ & - 2d_2 - 8D_1s_1 + 4d_1s_1 + 4d_1s_1^2 - 4d_1s_2] \\ & + \left(\frac{4}{n-1} - \frac{6}{n(n-1)}\right) \\ & \times [4(1 + 2s_1)(D_1 - G_1) + 8(G_2 - D_2) \\ & + (1 + 4s_1 + 4s_1^2 - 4s_2)(g_1 - d_1) \\ & + 2(g_1^2 - d_1^2 + 2(d_2 - g_2)] \\ & + \left(\frac{2}{n(n-1)}\right)[s_1 - 6s_2 + 8s_3 - 12s_1s_2 + 6s_1^2 \\ & + 4s_1^3 + 4G_1 - 8G_2 - g_1 - 2g_1^2 + 2g_2 + 8G_1s_1 \\ & - 4g_1s_1 - 4g_1s_1^2 + 4g_1s_2]. \end{aligned} \quad (18)$$

Finally, we define an ~95% critical value as

$$L_{new} = E(V_D) + 1.645\sqrt{\mathrm{Var}(V_D)}. \quad (19)$$

RESULTS AND DISCUSSION

To convince ourselves of the correctness of the above algebra and to demonstrate the inadequacy of $\mathrm{Var}(V_D)_{old}$ we used Monte Carlo simulations. Eleven artificial samples were constructed in the following way: The first data set containing 100 strains and 10 loci with five alleles at each locus was constructed from 96 strains of genotype

1 1 1 1 1 1 1 1 1 1

and one each of genotype

2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3
4 4 4 4 4 4 4 4 4 4
5 5 5 5 5 5 5 5 5 5.

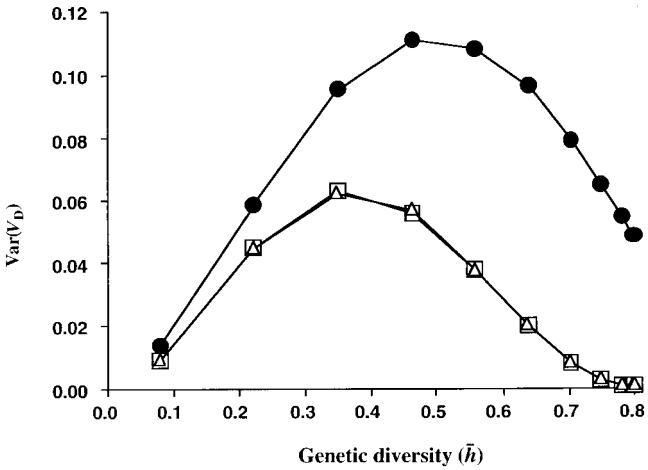

Figure 1.—Comparison between three methods of computing the variance of V_D,Var(V_D). Single random input matrices with 100 strains and 10 loci, each locus with the same genetic diversity (as described in the text), were analyzed and Var(V_D) computed according to Equation 6 (●), by resampling 10,000 times without replacement (△), or by using Equation 10 (□).

The second data set was made up of 88 strains of the major genotype and 3 strains of each of the minor genotypes and so on until a data set of maximum genetic diversity was reached consisting of 20 strains of each genotype. In this way we obtained artificial data sets with genetic diversities ranging from 0.078 to 0.8, which represent the range of genetic diversities to which the test developed by Brown *et al.* (1980) has been applied. The completely linked artificial data sets were then unlinked by one round of resampling without replacement.

For each sample, $\mathrm{Var}(V_D)_{old}$ and Var(V_D) were computed (using Equations 6 and 10, respectively). In addition, the randomization method suggested by Souza *et al.* (1992) was applied to each sample. That is, the alleles at each locus were shuffled randomly (resampling without replacement) and V_D calculated for each of 10,000 such shuffled samples. This allowed the calculation of the simulated sampling variance of V_D, $\mathrm{Var}(V_D)_{MC}$.

When $\mathrm{Var}(V_D)_{old}$ was compared with $\mathrm{Var}(V_D)_{MC}$, it was found that the two values diverged dramatically for input matrices of high genetic diversity (Figure 1). This causes similar divergence between true and estimated critical values (data not shown) and has implications for testing linkage equilibrium in bacterial populations that will be discussed later. Clearly, Equation 6 should not be used. No discrepancies were found between $\mathrm{Var}(V_D)_{MC}$ and the variance calculated with Equation 10 (see Figure 1).

The usefulness of (19) for hypothesis testing depends on whether the distribution of V_D is approximately normal under our null hypothesis of linkage equilibrium with replicates being produced by shuffling of alleles on haplotypes. For multilocus data sets there are three variables that may influence the shape of the distribu-

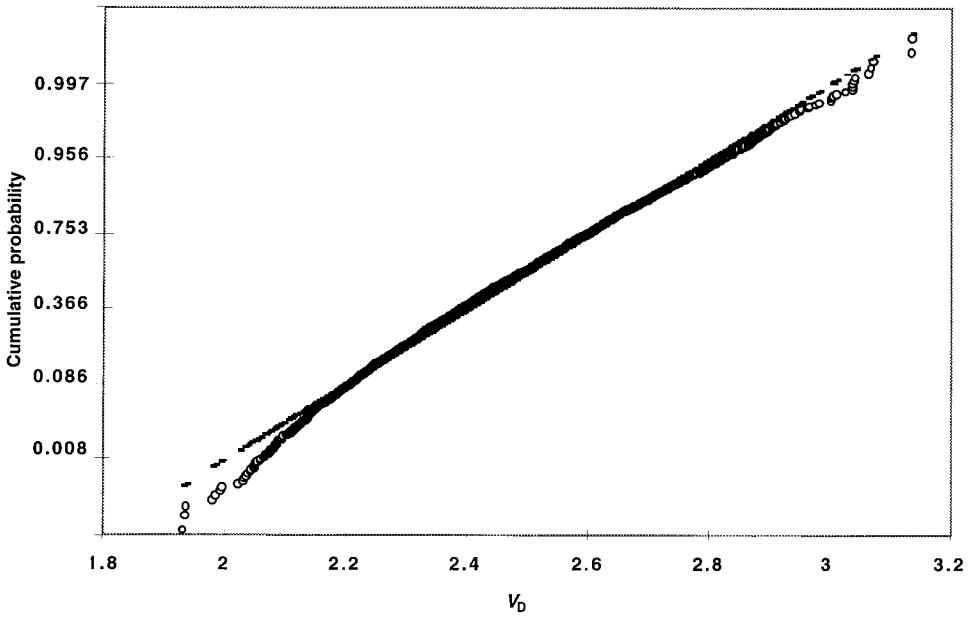

Figure 2.—Cumulative probability plot of 2500 resampled V_D values. The values expected if the distribution was normal (−) and those observed (○) diverge at both extremes of the distribution, although for testing the hypothesis of linkage equilibrium only the positive skew apparent in the high cumulative probability values is of interest. The resampled artificial input data set consisted of 100 strains and 10 loci, each with a genetic diversity of 0.558.

tion of V_D, the number of loci, the degree of diversity at each locus, and the number of strains. We investigated the effect of these three variables on the skewness of the distribution of V_D through Monte Carlo simulation by calculating $g1$ as a measure of skewness from sets of resampled V_D values,

$$g1 = \frac{m_3}{m_2^{3/2}}, \tag{20}$$

where m_3 and m_2 are the second and third moment of the distribution of V_D (Sokal and Rohlf 1981, p. 114). For a normal distribution $g1 = 0$; a positive $g1$ indicates skewness to the right, a negative $g1$, skewness to the left. We found that the distribution of V_D always had positive skewness, that is, at the upper extreme of the distribution, slightly more values lie beyond the normal critical values (Figure 2). This was not affected by the number of loci (data not shown). In contrast, the degree of genetic diversity at each locus had a strong effect on the shape of the distribution. On the whole, the greater the genetic diversity, the closer the distribution was to normality, but this relation was not linear with the strongest changes occurring at the extreme values of mean genetic diversity ($\bar{h}$; Figure 3). Sample size also had a strong effect on skewness. In general, the larger the sample, the closer the sampling distribution of V_D approached normality (Table 1).

Given that the distribution of V_D has positive skewness even for large samples, we investigated the effect of this deviation from normality on hypothesis testing. Data sets consisting of between 15 and 480 strains and 10 loci, each with genetic diversity of 0.444, were resampled to calculate the frequency with which V_D exceeded the critical values that would be obtained if the distribution of V_D was normal. Even for small data sets the discrepancy was slight. For instance, with 15 strains 6.69% of the resampled V_D values exceeded the 5% normal critical value (Table 1). For a sample of 480 strains the discrepancy between 5.13% and 5.0% was negligible. Note that the probabilities of exceeding the normal critical values were always slightly too large, as would be expected from the positive skewness of the distribution of V_D. For real data this means that whenever a sample has been diagnosed as being in linkage equilibrium, the same conclusion would be reached by Monte Carlo simulation. Further, the more time consuming it becomes to test the hypothesis of linkage equilibrium due to large sample size, the more useful our formula becomes. This is because the sampling distribution of V_D approaches normality for large samples.

Several recent reports of panmixis in bacteria have

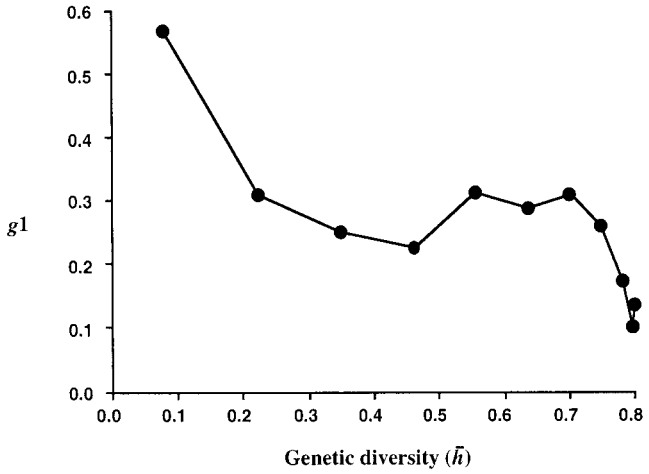

Figure 3.—Skewness of the distribution of V_D ($g1$) as a function of mean genetic diversity ($\bar{h}$). Single random input matrices with 100 strains and 10 loci, each locus with the same genetic diversity, were resampled 10,000 times without replacement.

TABLE 1

Relationship between skewness and the probability of exceeding normal critical values for various levels of significance

		Probability of exceeding normal critical values for α =				
n	$g1$	0.1	0.05	0.025	0.01	0.005
15	0.7229	0.1143	0.0669	0.0427	0.0256	0.0167
30	0.5125	0.1091	0.0649	0.0405	0.0213	0.0134
60	0.3683	0.1068	0.0609	0.0351	0.0164	0.0111
120	0.2965	0.1084	0.0603	0.0327	0.0156	0.0100
240	0.2061	0.1101	0.0564	0.0306	0.0137	0.0077
480	0.1445	0.1006	0.0513	0.0278	0.0111	0.0059

Single random input matrices were resampled without replacement 10,000 times and a V_D value computed each time. Subsequently, the frequency with which these V_D values exceeded the critical values obtained, assuming normality, was computed. n, number of strains; $g1$, measure of skewness of the distribution of V_D values obtained through resampling.

used the observed variance of pairwise differences (V_D) as a test statistic. Panmixis was concluded if the critical value of V_D was greater than the observed value of V_D (Maynard Smith *et al.* 1993; Bottomley *et al.* 1994; Duncan *et al.* 1994; Strain *et al.* 1995; Go *et al.* 1996). The original method to calculate the critical value was devised for plant populations, which are only moderately diverse [*e.g.*, $\bar{h}$ (*H. spontaneum*) = 0.145 (Brown *et al.* 1980)], compared to bacterial populations (*cf.* Table 1). In this study we showed by Monte Carlo simulation that high genetic diversity leads to an artificial inflation of $\mathrm{Var}(V_D)_{old}$ (Figure 1). This problem was overcome by rederiving $\mathrm{Var}(V_D)$ (Equation 10; Figure 1).

Bacterial populations: To test the usefulness of this derivation in the study of bacterial population genetics, we investigated published allozyme data for the ECOR collection of *E. coli* (Ochman and Selander 1984), which is a well-known example of a clonal population (Miller and Hartl 1986). In addition, data sets from *Bradyrhizobium japonicum*, *B. subtilis*, and *Rhizobium leguminosarum* were included in the analysis, because for these populations claims of linkage equilibrium have been based on incorrect formulas for the variance of V_D. Finally, an allozyme data set from *N. gonorrhoeae* was reexamined, as this taxon is considered a prime example of a sexual bacterial population (Maynard Smith *et al.* 1993; O'Rourke and Stevens 1993).

Generally we observed that bacterial populations are highly diverse ($\bar{h}$ = 0.311 to 0.691; Table 2) and that the genetic diversity varies strongly between loci (standard deviation = 0.178 to 0.304; Table 2). Further, the distribution of V_D displayed positive skewness in all cases, as observed in the simulations (Table 2).

E. coli: As expected from previous work (Miller and Hartl 1986), the electrophoretic types of the ECOR collection of *E. coli* are in linkage disequilibrium when the critical value obtained through the Monte Carlo process, L_{MC}, is compared to V_D($L_{MC} < V_D$; Table 2). Further, L_{new} (= 2.592) is a good estimator of L_{MC} (= 2.608), while L_{old} (= 2.985) not only overestimates the critical value of V_D, but would also lead to the spurious conclusion that *E. coli* is in linkage equilibrium as $L_{old} > V_D$ (Table 2).

B. japonicum: Bottomley *et al.* (1994) reported linkage equilibrium for a *B. japonicum* population represented by 17 electrophoretic types. This claim is clearly rejected by Monte Carlo simulation, which shows significant linkage for this population (L_{MC} = 2.593 < V_D = 3.985; Table 2). The same conclusion is reached by comparing L_{new} (= 2.557) with V_D. Surprisingly, V_D also exceeds L_{old}, on which the original claim of linkage equilibrium had been based. This discrepancy is resolved if L_{old} is calculated on the basis of the biased estimator

$$h_j^b = 1 - \sum_i p_{ij}^2 ,$$

rather than on the unbiased estimator (Equation 4) employed in this study. Using h_j^b, L_{old} = 3.996, which is slightly greater than V_D = 3.985. This result is due to the large difference between biased and unbiased estimators of the genetic diversity per locus in a sample consisting of only 17 ETs.

R. leguminosarum: Strain *et al.* (1995) obtained evidence of linkage disequilibrium in their U.K. population of *R. leguminosarum* by using Monte Carlo simulation, but H_0 was not rejected on the basis of L_{old}. We obtained the same result, reinforcing the inappropriateness of L_{old} for hypothesis testing. We further found that L_{new} (= 2.911) was again a good alternative to the lengthy calculations necessary for obtaining L_{MC} (= 2.967; Table 2) through simulation. Strain *et al.* (1995) also analyzed groups I + III + IX and I + III of their *R. leguminosarum* U.K. population and reported linkage equilibrium for both subpopulations. We found that H_0 is rejected for groups I + III + IX and I + III on the basis of L_{MC} and L_{new} (Table 2).

B. subtilis: Duncan *et al.* (1994) reported linkage equi-

TABLE 2

Assessment of multilocus structure in bacterial populations and comparison of old and new estimators of the critical values of V_D

Population	Sample size	Loci	$\bar{h}$	$g1$	V_D	$E(V_D)$	Estimators of the 95% critical value for V_D			Linkage detected?
							L_{MC}	L_{new}	L_{old}	
E. coli (Ochman and Selander 1984)	46 ETs	11	0.493 ± 0.225	0.440	2.954	2.245	2.608	2.592	2.985	Yes
B. japonicum (Bottomley *et al.* 1994)	17 ETs	13	0.691 ± 0.230	0.380	3.985	2.142	2.592	2.557	3.295	Yes
R. leguminosarum, UK population (Strain *et al.* 1995)	32 ETs	13	0.472 ± 0.275	0.449	3.414	2.332	2.731	2.695	3.250	Yes
R. leguminosarum, UK population, groups I + III + IX (Strain *et al.* 1995)	23 ETs	13	0.450 ± 0.261	0.569	3.467	2.392	2.967	2.911	3.497	Yes
R. leguminosarum, UK population, groups I + III (Strain *et al.* 1995)	18 ETs	13	0.400 ± 0.256	0.612	3.264	2.333	3.014	2.931	3.535	Yes
B. subtilis, groups B & D combined (Duncan *et al.* 1994)	50 ETs	13	0.491 ± 0.304	0.465	4.128	2.138	2.422	2.397	2.822	Yes
B. subtilis, group D (Duncan *et al.* 1994)	27 isolates	13	0.376 ± 0.302	0.531	2.953	1.957	2.387	2.347	2.795	Yes
B. subtilis, group B (Duncan *et al.* 1994)	28 isolates	13	0.354 ± 0.269	0.629	2.590	2.106	2.664	2.605	3.002	No
N. gonorrhoeae (O'Rourke and Stevens 1993)	228 isolates	9	0.311 ± 0.178	0.211	1.750	1.675	1.837	1.831	1.920	No

ET, electrophoretic type; $\bar{h}$, mean genetic diversity per locus ± standard deviation; V_D, observed variance of pairwise differences; $E(V_D)$, expected variance of pairwise differences in case of linkage equilibrium; L_{MC}, 95% critical value estimated by Monte Carlo simulation; L_{new}, 95% critical value as defined in Equation 19; L_{old}, 95% critical value as defined in Equation 15.

librium for the 50 electrophoretic types of *B. subtilis* contained in the B and D subdivisions of their sample. In contrast, we found that the combined electrophoretic types of groups B and D display strong linkage (Table 2) with V_D (= 4.128) far exceeding L_{MC} (= 2.422) and L_{new} (= 2.397). Group D on its own is also not in linkage equilibrium with L_{MC} and $L_{new} < V_D$, but note that as for *E. coli*, *R. leguminosarum*, and *B. japonicum*, application of L_{old} would lead to an inappropriate conclusion of linkage equilibrium. We further concluded on the basis of L_{MC} (= 2.664) and L_{new} (= 2.605) that group B is indeed in linkage equilibrium (Table 2).

N. gonorrhoeae: This group of bacteria is the best established example of a bacterial population in linkage equilibrium. An extensive allozyme data set comprising 228 isolates has been published and reported to be in linkage equilibrium (Maynard Smith *et al.* 1993; O'Rourke and Stevens 1993). Moreover, *N. gonorrhoeae* is naturally competent and frequently encounters different genotypes of the taxon due to the sexual habits of its host. As expected, we found that this population is in linkage equilibrium according to L_{MC} (= 1.837); L_{new} (= 1.831) gave the same result, further confirming the usefulness of this algebraic confidence limit (Table 2).

For all the bacterial populations tested, L_{MC} and L_{new} agreed well. This contrasted with the strong divergence of L_{old} from L_{MC}, which led to conflicting conclusions about the genetic structure of *E. coli*, *B. japonicum*, *R. leguminosarum*, and *B. subtilis*. Using computer simulations, Maynard Smith (1994) showed that a recombination rate only 20 times the rate of mutation was sufficient to unlink bacterial genomes. The detection of linkage disequilibrium in the soil-dwelling populations of *B. japonicum*, *R. leguminosarum*, and *B. subtilis* presented in this article indicates that the recombination rates in these groups are probably very low. This has also been found experimentally for *B. subtilis* (Roberts and Cohan 1995).

We conclude that past attempts to detect linkage disequilibrium in haploid multilocus data sets through the computation of a critical value for V_D were based on an erroneous formula for the variance of V_D. The correct formula for Var(V_D) communicated in this article forms the basis of a simple test of linkage. Furthermore, we find that V_D is approximately normally distributed (especially for large samples). Hence the algebraic test proposed here is a useful alternative to Monte Carlo simulation in cases where simulation is deemed too expensive or time consuming. A computer program written in FORTRAN77, which implements both the algebraic as well as the Monte Carlo test, can be obtained from B.H. upon request.

We thank J. Maynard Smith for first drawing our attention to the problem of testing linkage equilibrium from mismatch data and for helpful discussion. Thanks are also due to P. J. Bottomley for providing the Rhizobium allozyme data, and to T. S. Whittam and two anonymous reviewers for comments on the manuscript. This work was supported by grants from the Royal Society, Oxford University and the Biotechnology and Biological Sciences Research Council (United Kingdom).

LITERATURE CITED

Bottomley, P. J., H.-H. Cheng and S. R. Strain, 1994 Genetic structure and symbiotic characteristics of a *Bradyrhizobium* population recovered from a pasture soil. Appl. Environ. Microbiol. **60:** 1754–1761.

Brown, A. H. D., M. W. Feldman and E. Nevo, 1980 Multilocus structure of natural populations of *Hordeum spontaneum*. Genetics **96:** 523–536.

Duncan, K. E., N. Ferguson, K. Kimura, X. Zhou and C. Istock, 1994 Fine-scale genetic and phenotypic structure in natural populations of *Bacillus subtilis* and *Bacillus licheniformis*: implications for bacterial evolution and speciation. Evolution **48:** 2002–2025.

Go, M. F., V. Kapura, D. Y. Graham and J. M. Musser, 1996 Population genetic analysis of *Helicobacter pylori* by multilocus enzyme electrophoresis: extensive allelic diversity and recombinational population structure. J. Bacteriol. **178:** 3934–3938.

Guttman, D. S., and D. E. Dykhuizen, 1994 Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. Science **266:** 1380–1383.

Haubold, B., and P. B. Rainey, 1996 Genetic and ecotypic structure of a fluorescent *Pseudomonas* population. Mol. Ecol. **5:** 747–761.

Hudson, R. R., 1994 Analytical results concerning linkage disequilibrium in models with genetic transformation and conjugation. J. Evol. Biol. **7:** 535–548.

Istock, C. A., K. E. Duncan, N. Ferguson and X. Zhou, 1992 Sexuality in a natural population of bacteria: *Bacillus subtilis* challenges the clonal paradigm. Mol. Ecol. **1:** 95–103.

Maruyama, T., and M. Kimura, 1980 Genetic variability and effective population size when local extinction and recolonization of subpopulations are frequent. Proc. Natl. Acad. Sci. USA **77:** 6710–6714.

Maynard Smith, J., 1994 Estimating the minimum rate of genetic transformation in bacteria. J. Evol. Biol. **7:** 525–534.

Maynard Smith, J., N. H. Smith, C. G. Dowson and B. G. Spratt, 1993 How clonal are bacteria? Proc. Natl. Acad. Sci. USA **90:** 4384–4388.

Milkman, R., 1973 Electrophoretic variation in *Escherichia coli* from natural sources. Science **182:** 1024–1026.

Miller, R. D., and D. L. Hartl, 1986 Biotyping confirms a nearly clonal population structure in *Escherichia coli*. Evolution **40:** 1–12.

Ochman, H., and R. K. Selander, 1984 Standard reference strains of *Escherichia coli* from natural populations. J. Bacteriol. **157:** 690–693.

O'Rourke, M., and E. Stevens, 1993 Genetic structure of *Neisseria gonorrhoeae* populations: a non-clonal pathogen. J. Gen. Microbiol. **139:** 2603–2611.

Roberts, M. S., and F. M. Cohan, 1995 Recombination and migration rates in natural populations of *Bacillus subtilis* and *Bacillus mojavensis*. Evolution **49:** 1081–1094.

Selander, R. K., and B. R. Levin, 1980 Genetic diversity and structure in *Escherichia coli* populations. Science **210:** 545–547.

Sokal, R. R., and F. J. Rohlf, 1981 *Biometry*, Ed. 2. W. H. Freeman, New York.

Souza, V., T. T. Nguyen, R. R. Hudson, D. Piñero and R. E. Lenski, 1992 Hierarchical analysis of linkage disequilibrium in *Rhizobium* populations: evidence for sex? Proc. Natl. Acad. Sci. USA **89:** 8389–8393.

Strain, S. R., T. S. Whittam and P. J. Bottomley, 1995 Analysis of genetic structure in soil populations of *Rhizobium leguminosarum* recovered from the USA and the UK. Mol. Ecol. **4:** 105–114.

Whittam, T. S., H. Ochman and R. K. Selander, 1983 Multilocus genetic structure in natural populations of *Escherichia coli*. Proc. Natl. Acad. Sci. USA **80:** 1751–1755.

Wise, M. G., L. J. Shimkets and J. V. McArthur, 1995 Genetic structure of a lotic population of *Burkholderia* (*Pseudomonas*) *cepacia*. Appl. Environ. Microbiol. **61:** 1791–1798.

Communicating editor: P. L. Foster