# MULTILOCUS STRUCTURE OF NATURAL POPULATIONS OF *HORDEUM SPONTANEUM**

A. H. D. BROWN,[1] M. W. FELDMAN[2] AND E. NEVO[3]

[1] *CSIRO, Division of Plant Industry, Canberra, A.C.T. 2601, Australia*
[2] *Department of Biological Sciences, Stanford University, Palo Alto, California 94305*
[3] *Institute of Evolution, Haifa University, Israel*

Manuscript received January 29, 1980
Revised copy received August 12, 1980

## ABSTRACT

The association of alleles among different loci was studied in natural populations of *Hordeum spontaneum*, the evolutionary progenitor of cultivated barley. The variance of the number of heterozygous loci in two randomly chosen gametes affords a useful measure of such association. The behavior of this statistic in several particular models is described. Generally, linkage (gametic phase) disequilibrium tends to increase the variance above the value expected under complete independence. This increase is greatest when disequilibria are such as to maximize the sum of squares of the two-locus gametic frequencies.——When data on several loci per individual are available, the observed variance may be tested for its agreement with that expected under the hypothesis of complete interlocus independence, using the sampling theory of this model. When applied to allozyme data from 26 polymorphic populations of wild barley, this test demonstrated the presence of geographically widespread multilocus organization. On average, the variance was 80% higher than expected under random association. Gametic frequencies for four esterase loci in both of these populations of wild barley and two composite crosses of cultivated barley were analyzed. Most generations of the composites showed less multilocus structure, as measured by the indices of association, than the wild populations.

THE concept of linkage or gametic disequilibrium, or the association of particular alleles among loci, is central to modern population genetics (see WEIR 1979; HEDRICK, JAIN and HOLDEN 1978; KARLIN 1975 for recent reviews). Several measures of gametic disequilibrium have been used. The most common is:

$$D_{ik} = g\,(A_iB_k) - p(A_i)\;p(B_k) \qquad (1)$$

where $D_{ik}$ is the difference between the observed frequency of the gamete carrying both alleles $A_iB_k$ [$g\ (A_iB_k)$] and its expected frequency assuming no association [$p\ (A_i)\ p\ (B_k)$]. This and related measures are relatively manageable when there are few alleles at a few loci. However, with multiple alleles and many loci, the number of measures quickly increases beyond comprehension. It would be desirable to have a set of summary statistics that, in some defined

* Research supported in part by National Science Foundation grant DEB7705742, Public Health Service grants GM 10452 and GM 28016 and a United States-Israel Binational Science Foundation grant.

sense, describes the extent of multilocus structure, at the risk of subsuming information on particular allelic combinations. This is analogous to summarizing genetic diversity at single loci by such measures as the mean number of alleles per locus, or the mean panmictic heterozygosity per locus [an analog of SIMPSON's (1949) index of species diversity]. Summary statistics could be useful for comparing various kinds of species. Population genetic data on several loci of the same individual are now commonly collected, using electrophoresis, yet multilocus patterns are rarely studied.

Following a suggestion of SVED (1968), we propose to use multilocus measures, related to the single-locus SIMPSON index, to measure multilocus association. These measures are based on the observed distribution of the random variable, *the number of heterozygous loci in two randomly chosen gametes*, and in particular, on the lower central moments of this distribution. The statistic of most use is the variance of this distribution. This quantity has received some theoretical attention (SVED 1968; HILL 1975; FRANKLIN 1977; AVERY and HILL 1979). More importantly, however, the variance has a biological appeal. For a given mean heterozygosity, a population with a high variance will, on outbreeding, produce more multiply heterozygous individuals and more multiply homozygous individuals than if the variance were low. On inbreeding, such a population produces few strains with greater genetic differentiation between them.

In this paper, we shall first develop the underlying theory of our measures and then apply them to studies of genetic organization in two composite populations of barley, *Hordeum vulgare* (CLEGG, ALLARD and KAHLER 1972) and Israeli populations of wild barley, *H. spontaneum* (NEVO *et al.* 1979).

## THEORY OF MEASURES OF MULTILOCUS STRUCTURE

Consider an infinite population of gametes with genotype known at each of $m$ loci. The random variable $K$ is defined as the number of loci that are different (heterozygous) when two such random gametes are compared at these loci ($K = 0, 1, \ldots, m$). The distribution of this random variable, $f(K)$, will depend on the degree of polymorphism at each of the $m$ loci, as well as on any correlations in allelic variants among loci over gametes. In this section, the relations between the lower moments of the distribution of $K$, the allele frequencies and conventional parameters of gametic disequilibrium [*e.g.*, formula (1) above] are given in several cases. The indices of multilocus structure are then defined, and a test of the null hypothesis that the observed variance equals that expected under independent association of alleles is outlined.

*Interlocus independence of alleles:* Suppose that no correlation exists and that the alleles at each locus are entirely independent in occurrence. Let $p_{ji}$ denote the population frequency of the $i^{\text{th}}$ allele at the $j^{\text{th}}$ locus, and let $h_j$ $(= 1 - \sum_i p_{ji}^2)$ denote the genic diversity (the panmictic heterozygosity) at the $j^{\text{th}}$ locus. Under the assumption of locus independence, the random variable $K$ is distributed as the sum of $m$ independent Bernoulli distributions with moment-generating function:

$$G(\theta) = \prod_{j=1}^{m} (h_j e^{\theta} + 1 - h_j) \ . \qquad (2)$$

Differentiating this generating function with respect to $\theta$ and setting $\theta = 0$ yields $E(K) = \Sigma h_j$, and the following central moments:

$$\sigma^2_K = E[K - E(K)]^2 = \Sigma h_j - \Sigma h^2_j \qquad (3)$$

$$E[K - E(K)]^3 = \Sigma h_j - 3\Sigma h^2_j + 2\Sigma h^3_j \qquad (4)$$

$$E[K - E(K)]^4 = \Sigma h_j - 7\Sigma h^2_j + 12\Sigma h^3_j - 6\Sigma h^4_j + 3[\Sigma h_j - \Sigma h^2_j]^2 \qquad (5)$$

where all summations are over $j = 1,2, \ldots, m$. If the $\{h_j\}$ are all equal to $h$, formula (5) reduces to the form of KENDALL and STUART's (1969) formula 5.5.

*Two loci, multiple alleles, general case:* Let the frequency of the gamete with with the $i^{\text{th}}$ allele at the $A$ locus and the $k^{\text{th}}$ allele at the $B$ locus be:

$$g_{ik} = p_{1i}\, p_{2k} + D_{ik} \ . \qquad (6)$$

The frequency distribution for the number of loci that differ in their allelic constitution in a pair of gametes, $f(K)$; $K = 0,1,2,$ is

$$f(0) = \sum_{ik}\sum g^2_{ik} \qquad (7)$$

$$f(1) = \sum_i p^2_{1i} + \sum_k p^2_{2k} - 2 \sum_{ik}\sum g^2_{ik}$$

$$f(2) = 1 - \sum_i p^2_{1i} - \sum_k p^2_{2k} + \sum_{ik}\sum g^2_{ik} \ .$$

From this it follows that the mean and variance are:

$$E(K) = 2 - \sum_i p^2_{1i} - \sum_k p^2_{2k} = h_1 + h_2 \qquad (8)$$

$$\sigma^2_K = h_1 + h_2 - h^2_1 - h^2_2 + 2\, \Sigma\Sigma\, g^2_{ik} - 2\,(1-h_1)(1-h_2)$$

$$= (2-h_1-h_2)(h_1+h_2-1) + 2 \sum_{ik}\sum g^2_{ik} \qquad (9)$$

$$= h_1 + h_2 - h^2_1 - h^2_2 + 4 \sum_{ik}\sum p_{1i}p_{2k}D_{ik} + 2 \sum_{ik}\sum D^2_{ik} \ . \qquad (10)$$

The deviation of the variance of the number of heterozygous loci ($\sigma^2_K$) from its value (formula 3), assuming the loci are independent, is given by the two terms in expression (10) that involve linkage disequilibrium. Given fixed values for the allele frequencies $\{p_{1i};\ i = 1,2, \ldots, r$ and $p_{2k};\ k = 1,2, \ldots, s\}$, where there are $r$ alleles at the first locus and $s$ at the second locus, the variance will depend on the coefficients of disequilibrium $\{D_{ik}\}$. This conditional variance (given the allele frequencies) can be differentiated with respect to the $\{D_{ik}\}$, subject to the $(r + s)$ side-conditions:

$$\sum_i D_{ik} = \sum_k D_{ik} = 0.$$

The solution for the one stationary point is:

$$D_{ik} = -\ (rp_{1i} - 1)\ (sp_{2k} - 1)/rs \ . \qquad (11)$$

This solution is permissible only if

$$rsp_{1i}, rsp_{2k} \geq rp_{1i} + sp_{2k} - 1 \geq 0, rs(p_{1i} + p_{2k} - 1) \ .$$

The point is a minimum because all second differentials are positive. At this point, the variance is

$$\min \{\sigma^2_K\} = h_1 + h_2 - h^2_1 - h^2_2 - 2\,[1-h_1-1/r]\;[1-h_2-1/s] \ . \qquad (12)$$

The last term is either positive or zero for completely "even" allele frequency distributions. In general, therefore, as disequilibria depart from the value (11), the conditional variance increases. Comparing (10) and (12), the minimum point coincides with zero disequilibrium only when at least one of the loci has an "even" frequency distribution (*i.e.*, $p_{1i} = 1/r$, for all $i$).

Unfortunately, the maximum value of the conditional variance is harder to define, as it depends on the side conditions that $g_{ik} \geq 0$, for all $i$ and $k$, and $\sum\sum_{ik} g_{ik} = 1$. It is clear from equation (9) that, for fixed $\{h_j\}$, the variance is a maximum when the disequilibria are such as to maximize the sum of squares of the two-locus gametic frequencies. This property is comparable to the behavior of the single-locus diversity measure $\{h_j\}$ under various allelic frequency distributions. For a fixed number of alleles, the diversity is a maximum in "even" distributions, when the sum of squares of allelic frequencies is minimized.

*Three loci, two alleles per locus:* Our aim in this section is to show how first-order disequilibria over several loci and the second-order disequilibrium contribute to the central moments of the distribution $f(K)$. For convenience, the notation in the three-locus two-allele case is varied to reduce the subscripts, and the frequency of the allele $A_1$, at the $A$ locus is denoted as $p$, that of $A_2$ is $q$ ($= 1$-$p$), of $B_1$ is $u$, of $B_2$ is $v$ ($= 1$-$u$) and $C_1$ is $x$ and $C_2$ is $y$ ($= 1$-$x$), the disequilibrium between the $AB$ loci as $D_1$, between $AC$ as $D_2$ and $BC$ and $D_3$, and the second-order disequilibrium as $T$. The eight three-locus gametic frequencies are functions of these seven parameters (specifically in this notation, in Table V of BROWN 1975). For example, $P(A_1B_1C_1) = pux + pD_3 + uD_2 + xD_1 + T$. The genic diversities for the $A$, $B$ and $C$ loci are $h_1$, $h_2$ and $h_3$, respectively. The frequency distribution for the number of heterozygous loci in two random uniting gametes,

$$\{f(K); K = 0,1,2,3\} \ ,$$

is computed in terms of these seven parameters. The following are the moments of this distribution:

$$E(K) = \sum_{j=1}^{3} h_j$$

$$\begin{aligned}\sigma^2_K = \sum_{j=1}^{3} h_j(1-h_j) + 4D_1(q-p)(v-u) + 4D_2(q-p)(y-x)\\ + 4D_3(v-u)(y-x) + 8\sum_{j=1}^{3} D^2_j \qquad (13)\end{aligned}$$

$$\begin{aligned}E[K-E(K)]^3 = \sum_{j=1}^{3} h_j(1-h_j)(1-2h_j)\\ - 12(q-p)(v-u)[2D_2D_3 + D_1(1+h_1+h_2)]\\ - 12(q-p)(y-x)[2D_1D_3 + D_2(1+h_1+h_3)]\end{aligned}$$

$$\begin{aligned}&-12(v-u)(y-x)[2D_1D_2+D_3(1+h_2+h_3)]\\&+24D_1^2(1-h_1-h_2)+24D_2^2(1-h_1-h_3)+24D_3^2(1-h_2-h_3)\\&+48(y-x)D_1T+48(v-u)D_2T+48(q-p)D_3T\\&+12T(q-p)(y-x)(v-u)-48T^2\ .\end{aligned}\tag{14}$$

Thus, the two-locus associations between all possible pairs of $m$ loci contribute additively to the variance of $K$. In general:

$$\sigma_K^2=\overset{m}{\Sigma}h_j-\overset{m}{\Sigma}h_j^2+2\overset{m}{\underset{j}{\Sigma}}\overset{m}{\underset{l>j}{\Sigma}}\underset{i}{\Sigma}\underset{k}{\Sigma}[2p_{ji}p_{lk}D_{ik}^{jl}+(D_{ik}^{jl})^2]\tag{15}$$

where $D_{ik}^{jl}$ is the disequilibrium between the $i^{\text{th}}$ allele at the $j^{\text{th}}$ locus and the $k^{\text{th}}$ allele at the $l^{\text{th}}$ locus. SVED's (1968) equation (4) and AVERY and HILL's (1979) equation (10) are both diallelic cases of (15). An alternative expression for (15), which is in the format of (9), is:

$$\sigma_K^2=(m-\Sigma h_j)(\Sigma h_j+1-m)+2\underset{j}{\Sigma}\underset{l>j}{\Sigma}\underset{i}{\Sigma}\underset{k}{\Sigma}(g_{ik}^{jl})^2\tag{16}$$

where $g_{ik}^{jl}$ is the frequency of the two-locus gamete with the $i^{\text{th}}$ allele a the $j^{\text{th}}$ locus and the $k^{\text{th}}$ allele at the $l^{\text{th}}$ locus. The variance ($\sigma_K^2$) is independent of higher-order associations and is cumulative over locus pairs. It therefore tends to increase with "more" disequilibria within a pair and over pairs. The third central moment (14) is a complex function of both second- and third-order association. If, however, there is no first-order association, so that all the $D_i$ are zero, formula (14) reduces considerably. Thus, the third moment of the distribution could be a useful index of more complex disequilibria when two-locus disequilibria are not evident, especially when allele frequency distributions are "even."

*Multiple loci, absolute and complete association:* As the number of loci are increased, if they are independent, the variance of $K$ increases (15). It would be desirable for any index of association to have a standard that did not increase with increasing numbers of loci scored. Furthermore, such an index should ideally be independent of single-locus allele frequency distributions. One possibility is to divide the variance (15) by its expected value, assuming independence and the same single-locus diversities (3). We now consider some particular cases of intense association to see what standardization is achieved by such a division.

CLEGG *et al.* (1976) referred to the distinction between absolute association and complete association. For absolute multilocus association, each allele at each locus is uniquely associated with a distinctive allele at all other loci (*e.g.*, $A_1B_1C_1D_2\ldots$, $A_2B_2C_2D_1\ldots$). Complete association however signifies only that some combinations are completely lacking (*e.g.*, $A_1B_1$, $A_1B_2$, $A_2B_1$ present; $A_2B_2$ absent).

Consider $m$ polymorphic loci at each of which the alleles are absolutely associated. Therefore, $p_{ji}$, the frequency of the $i^{\text{th}}$ allele at the $j^{\text{th}}$ locus, equals $p_i$ for all *m loci*. Further:

$$\underset{j}{\Sigma}h_j=m(1-\underset{i}{\Sigma}p_i^2)$$

$$\underset{j}{\Sigma}h_j^2=m(1-\underset{i}{\Sigma}p_i^2)^2\ .$$

The distribution, $\{f(K); K = 0,1, \ldots, m\}$, of the number of heterozygous comparisons in two random gametes is:

$$f(0) = \sum_i p_i^2; f(m) = 1 - \sum_i p_i^2 \text{ and } f(K) = 0 \text{ for } K \neq 0, m.$$

The variance of this distribution is:

$$\sigma_K^2 = m^2 \sum_i p_i^2 [1 - \sum_i p_i^2] \ .$$

Then the ratio of the variance, to that expected for complete independence (3) is:

$$\sigma_K^2 / [\sum_j h_j - \sum_j h_j^2] = m \ . \qquad (17)$$

This ratio is independent of the allele frequencies, but is directly proportional to the number of loci scored. Thus, in the case of absolute association, standardization of the effect of locus number is not achieved by the division.

The case of complete association is much more complex because it covers all kinds of absences, starting with the absence of just one single gametic type. The question is whether the ratio examined in (17) would be expected to increase with increasing number of loci ($m$), but where the proportion of completely associated loci to total loci scored remains constant.

One simple example of this problem is as follows: Suppose that only two loci are studied and each has two alleles. Let the gamete frequency of $A_1B_1$ be 1–2$p$, and those of $A_1B_2$ and $A_2B_1$ each be $p$. (When $p = 0.5$, we have the simplest case of absolute association.) Then, using (10), the ratio is:

$$\frac{\sigma_K^2}{\sum_j h_j - \sum_j h_j^2} = \frac{1 - 4p + 8p^2 - 4p^3}{(1-p)[1-2p(1-p)]} \ . \qquad (18)$$

Now, suppose that gametes are scored at two additional loci where frequency of $C_1D_1$ is again 1–2$p$, and those of $C_1D_2$ and $C_2D_1$ are $p$. Further suppose that alleles at loci $C$ and $D$ are independent of those at $A$ and $B$, *i.e.*, the frequency of the gamete $A_1B_1C_1D_1$ is $(1–2p)^2$. The ratio, when computed, has the same value as (18) and thus is independent of the number of sets of similarly associated loci. [When $p = 0.5$, the ratio (18) equals 2, in agreement with (17).] Clearly, however, the ratio is not generally independent of either the number of loci scored or the single-locus allele frequencies. Yet, in certain circumstances (namely independence of loci and/or absolute association), the ratio may become independent of either or both of these. Generally, the ratio achieves only limited standardization. It may prove impossible to design a better statistic for association intensity for comparing the results from different studies that are based on radically different numbers of loci or on loci with widely differing levels of polymorphism..

*Indices of intensity of multilocus structure:* With the above treatment and qualifications in mind, we propose a series of indices of multilocus structure based on the observed moments of the distribution of the number of heterozygous comparisons. Let $m(i)$ denote the values of the $i^{th}$ central moment computed from the observed distribution, and let $\mu(i)$ denote the expected value of the $i^{\text{th}}$ central moment, given the single-locus diversities $(h_j)$ and the assumption of independence of loci [computed from (3), (4) and (5) above]. Then we define:

$$X(i) = [m(i)/\mu(i)] - 1 \qquad (i = 2,3, \ldots) \tag{19}$$

as measures of multilocus structure. [For $i = 1$, $X(1)$ equals zero.] For $i = 2$,

$$X(2) = s_K^2/(\Sigma h_j - \Sigma h_j^2) - 1 \tag{20}$$

where $m(2) = s_K^2$. For $i = 3$,

$$X(3) = m(3)/(\Sigma h_j - 3\Sigma h_j^2 + 2\Sigma h_j^3) - 1. \tag{21}$$

The division by $\mu(i)$ renders the $X$ measures less dependent on the values of $h_j$, and subtracting one gives expected values of the $X(i)$ of zero for all $i$ under independence.

*Estimation:* Two different experimental situations may be envisaged. In the first case (I) of a predominantly inbreeding species (such as *Hordeum spontaneum*), nearly all plants scored are homozygous for all marker loci. The data then are the observed frequencies of the $m$-locus gametic types. The second case (II) is in random-mating populations, in which each locus is in Hardy-Weinberg equilibrium.

The moments of the distribution can be estimated by two different methods. One method (A) is based on formula (15) and conceivably similar expressions not derived here for the higher moments. From a random sample of $n$ gametes (I) or $n$ zygotes (II), the values of $\{h_j\}$ for all loci $\{p_{ji}\}$ for all alleles and $\{D_{ik}^{jl}\}$ for all alleles at all possible pairs of loci are estimated. Then, the single statistic $s_K^2$, which is an estimate of the population value $\sigma_K^2$, is computed from formula (15). It is known from the estimation theory of the $\{h_j\}$ (SIMPSON 1949) and the $\{D_{ik}^{jl}\}$ (HILL 1974) that the sample estimates of those parameters have a bias factor of the order of $n^{-1}$. Therefore, $s_K^2$ has a bias of similar order of magnitude.

The second method (B) is by forming an empirical distribution of the number of heterozygous comparisons and computing the moments of this distribution. In the case of $m$-locus gametic data (case I), every gamete in the sample is compared with itself and every other gamete in turn, and the number of heterozygous loci is recorded. Thus, there are $n^2$ comparisons that form a Punnett square We have found that this procedure gives an estimate of $s_K^2$ numerically identical to that of method A above; in addition, it permits estimates of the higher moments to be readily computed.

In the case of multilocus diploid zygotes (II above), it is possible to record directly the number of heterozygous loci per individual to obtain the empirical distribution. Any departure from exact panmixia would, however, lead to estimates numerically different from method A. The bias and efficiency of this method have yet to be investigated.

*Hypothesis testing:* We now consider the null hypothesis that allelic distributions among loci are independent. The observed variance of the number of heterozygous gametic comparisons (case I) or loci (case II) is denoted as $s^2_K$, which is an estimate of the population value ($\sigma^2_K$). The null hypothesis is:

$$H_0\colon \sigma^2_K = \Sigma h_j - \Sigma h^2_j \ .$$

Assuming that $H_0$ is true, the sampling variance of the observed variance ($s^2_K$) is obtainable approximately (to the order of $n^{-1}$) using KENDALL and STUART's formula (12.14):

$$\mathrm{var}[s^2_K | H_0 \text{ is true}] = \{\mu(4) - [\mu(2)]^2\}/n \ .$$

From results (3) and (5) above

$$\mathrm{var}[s^2_K \mid H_0 \text{ is true}] = \{\Sigma h_j - 7\Sigma h^2_j + 12\Sigma h^3_j - 6\Sigma h^4_j - 2[\Sigma h_j - \Sigma h^2_j]^2\}/n \ , \qquad (22)$$

where the parametric values of the $\{h_j\}$ are replaced by their estimates. This is also the variance in case II, assuming independence among the loci. Using this variance and assuming the sampling distribution of $s^2_K$ approximates normality, the upper 95% confidence limit for $s^2_K$ is:

$$L \cong \Sigma h_j - \Sigma h^2_j + 2\,\{\mathrm{var}[s^2_K \mid H_0 \text{ is true}]\}^{1/2} \ . \qquad (23)$$

Thus, if the observed $s^2_K$ exceeds $L$, the null hypothesis of independence at the level of locus pairs is rejected. At this time, it is not clear whether the sampling distribution of $s^2_K$ is normal or whether another approximation would be appropriate. This is a problem worth pursuing.

## RESULTS

*Analysis of allozyme data from populations of* H. spontaneum: The data for this analysis came from a recent survey of the allozyme variation in 28 natural populations of wild diploid barley (*Hordeum spontaneum*) from seven distinct regions in Israel. Single seedlings from a total of 1,179 spikes were examined. The isozyme techniques and single-locus diversity analysis are described by BROWN, ZOHARY and NEVO (1978), whereas details of the populations and allozyme-environment relationships are in NEVO *et al.* (1979). Of the 28 loci screened, the multilocus genotypes of 20 of these were used in the analysis: *Acph-1*, *-2 -3*; *Adh-1*, *-2*; *Est-1*, *-2*, *-4*, *-5*; *Gdh*, *Got-1*, *-2*; *Mdh-1*, *-2*, *Nadhd-1*, *Pept-2*, *Pgi*, *Pgm*, *6Pgd-2*, *To-2*. The loci deleted and reasons were as follows: *Ald*, *Pepc*, *To-1* for complete invariance; *Cat*, *Nadhd-2*, *6Pgd-1* for near invariance; *Pept-1* for insufficient data and *Gp* for potentially complex inheritance; *Pept-1* for insufficient data and *Gp* for potentially complex inheritance. The aim of the laboratory routine was to assay all loci on all individuals. However, for various technical and strategic reasons, the data are incomplete. Among the 20 loci, the locus most deficient in scores was *Est-5*. Another problem with the data was that 43 seedlings were heterozygous at one or more loci. However, all the corresponding spikes had been resampled for estimation of outcrossing rate (BROWN, ZOHARY and NEVO 1978). Therefore, such scores could be changed on editing into those of the first detected fully homozygous sib of the original heterozygous seedling. Since outcrossing (overall average proportion was 0.016) and heterozygosity (> 96% of lines were fully homozygous) was relatively infrequent except at Talpiyyot, these corrections had minimal effect on the estimates.

The above multilocus analysis assumes that the data are complete. To overcome the problem of incomplete data, three computations were made: (1) Statistics were estimated first with the raw data, where heterozygous scores were regarded as missing data. Every gamete in the sample was compared with itself and every other gamete in turn and the number of definite heterozygous loci per combination recorded to form the underlying distribution, $f(K)$.

(2) Under the null hypothesis, the statistic $s^2_K$ is generally biased down when $h_j < 0.5$ and the data are incomplete. A correction factor can be derived from formulas (2) and (3) above, in which $h_j$ is replaced by $n^2_j h_j / n^2$, where $n_j$ is the number of homozygotes scored at the $j^{\text{th}}$ locus, $n$ is the total number of gametes and $h_j$ is the diversity at the $j^{\text{th}}$ locus. This correction factor is:

$$\sum_j (1-n^2_j/n^2)\ h_j\ [1-(1+n^2_j/n^2)\ h_j]\ . \tag{24}$$

The observed variance, adjusted for incomplete comparisons, is computed by adding this factor to the observed $s^2_K$ from computation (1).

(3) An edited subset of the data was prepared in which heterozygous scores were converted to homozygotes, using the above sib method; plants with assays that lacked scores for loci known to be locally polymorphic were omitted.

Table 1 summarizes the genetic statistics of each population. The two completely monomorphic populations [Mt. Hermon (1), Mt. Meron (9)] are omitted. The number of variable loci $m'$, number of individuals included ($n$), equivalent to the number of independent gametes sampled, and mean genetic diversity per variable locus ($h$) are listed with the multilocus measures. The statistic $\mu(2)$ is the expected variance of the number of heterozygous loci, assuming complete independence, computed from (3), and $L$ is the 95% upper confidence limit. This value is for comparison with ${}^{\sim 2}_{K}$, the observed variance. Three estimates of $s^2_K$ are given corresponding to the three computations detailed above. The values of the third estimate concurred with the adjusted estimates ($s^2_K$, column 2). The overall means indicate that the correction factor (24) underestimated the effect of missing data, because it assumes the null hypothesis of independence. Finally, estimates of $X(2)$, which are based on the adjusted values of $s^2_K$, $X(3)$ and $X(4)$, are listed.

All 26 polymorphic populations gave an adjusted value of $s^2_K$ that exceeded $\mu(2)$. Statistically significant increases ($s^2_K$ exceeding $L$) were found in 17 populations for original $s^2_K$ and 20 for adjusted values. Thus, there is evidence of widespread gametic phase disequilibrium in these populations. This conclusion was supported by the estimates of $s^2_K$ based on the subset of complete data (method 3).

The values of $X(2)$, the measure of intensity of multilocus association, ranged from 0.13 to 2.79: Values in excess of 1.0 indicate that the variance in heterozygous loci due to multilocus structure is more than double that due to polymor-

TABLE 1

*Estimates of multilocus genetic parameters in 26 polymorphic populations of* Hordeum spontaneum

| | | | | | | $s^2_K$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Population | $m'$ | $n$ | $\bar{h}$ | $\mu(2)$ | $L$ | (1) | (2) | (3) | $X(2)$ | $X(3)$ | $X(4)$ |
| 2 Shifon | 6 | 50 | 0.070 | 0.85 | 1.18 | 0.84 | 0.97 | 1.10 | 0.13 | 0.3 | —0.2 |
| 3 Afiq | 8 | 32 | 0.194 | 1.68 | 2.48* | 3.61 | 3.76 | 4.04 | 1.23 | 34.4‡ | 2.5 |
| 4 Tel Hay | 6 | 50 | 0.153 | 1.49 | 2.04* | 2.78 | 2.98 | 3.09 | 1.00 | 55.5‡ | 1.3 |
| 5 Rosh Pinna | 8 | 46 | 0.125 | 1.55 | 2.17* | 3.76 | 4.07 | 5.12 | 1.62 | 9.2 | 2.8 |
| 6 Gadot | 11 | 49 | 0.246 | 2.32 | 3.24(*) | 3.01 | 3.55 | 3.90 | 0.53 | —11.1 | 0.6 |
| 7 Tabigha | 10 | 47 | 0.210 | 2.23 | 3.13* | 3.55 | 3.77 | 3.26 | 0.69 | — 3.5 | 1.1 |
| 8 Zefat | 7 | 50 | 0.135 | 1.35 | 1.86(*) | 1.77 | 2.25 | 2.14 | 0.67 | 13.4 | 0.6 |
| 10 Maalot | 7 | 51 | 0.076 | 1.15 | 1.61* | 1.90 | 2.28 | 2.54 | 0.98 | 4.7 | 2.2 |
| 11 Damon | 7 | 53 | 0.120 | 1.44 | 1.97* | 2.94 | 3.03 | 3.77 | 1.11 | 1.8 | 1.5 |
| 12 Shechem | 9 | 30 | 0.115 | 1.46 | 2.20* | 2.62 | 2.65 | 3.32 | 0.82 | 4.0 | 2.0 |
| 13 Bar Giyyora | 8 | 53 | 0.141 | 1.50 | 2.06(*) | 2.02 | 2.22 | 2.53 | 0.48 | 2.9 | 0.7 |
| 14 Talpiyyot | 10 | 32 | 0.205 | 2.10 | 3.11* | 3.43 | 3.85 | 4.76 | 0.84 | 10.8 | 0.9 |
| 15 Eyzariya | 10 | 30 | 0.223 | 2.01 | 3.03 | 2.61 | 2.59 | 2.85 | 0.29 | —55.1 | 0.8 |
| 16 Tel Shoqet | 9 | 32 | 0.155 | 1.65 | 2.45 | 2.28 | 2.34 | 2.10 | 0.42 | — 2.3 | 0.5 |
| 17 Bor Mashash | 8 | 52 | 0.063 | 1.01 | 1.41* | 3.82 | 3.82 | 3 82 | 2.79 | 15.5 | 17.0 |
| 18 Revivim | 6 | 31 | 0.073 | 0.97 | 1.46* | 2.42 | 2.55 | 3.14 | 1.64 | 12.6 | 8.1 |
| 19 Yeroham | 7 | 35 | 0.119 | 1.26 | 1.84* | 2.35 | 2.42 | 2.15 | 0.92 | 4.8 | 1.4 |
| 20 Sede Boqer | 9 | 32 | 0.168 | 1.78 | 2.65 | 2.20 | 2.46 | 2.58 | 0.38 | 0.6 | 0.2 |
| 21 Bet Shean | 7 | 41 | 0.142 | 1.27 | 1.80 | 1.51 | 1.60 | 1.42 | 0.27 | —11.8† | 0.3 |
| 22 Mechola | 8 | 47 | 0.117 | 1.32 | 1.86 | 1.73 | 1.75 | 1.76 | 0.32 | 1.2 | 0.7 |
| 23 Wadi Qilt | 9 | 45 | 0.154 | 1.72 | 2.42* | 3.21 | 3.42 | 3.52 | 0.99 | 4.4 | 2.1 |
| 24 Akhziv | 9 | 49 | 0.200 | 1.99 | 2.75* | 4.18 | 4.62 | 5.73 | 1.32 | —33.0 | 1.8 |
| 25 Atlit | 11 | 49 | 0.193 | 1.99 | 2.77* | 3.32 | 3.32 | 3.38 | 0.67 | —13.8 | 1.8 |
| 26 Caesarea | 9 | 48 | 0.080 | 1.16 | 1.63* | 1.72 | 1.91 | 1.70 | 0.65 | 2.8 | 1.4 |
| 27 Herzliyya | 7 | 50 | 0.109 | 1.41 | 1.95* | 2.22 | 2.22 | 2.22 | 0.58 | 1.2 | 1.1 |
| 28 Ashqelon | 9 | 45 | 0.186 | 1.82 | 2.56* | 2.64 | 2.69 | 2.74 | 0.48 | — 7.1 | 0.8 |
| Average | 8.3 | 43 | 0.145 | 1.56 | 2.22 | 2.63 | 2.81 | 3.02 | 0.80 | 2.3 | 1.6 |

Symbols: $m'$ = number of polymorphic loci; $n$ = number of plants sampled; $\bar{h}$ = mean single-locus diversity = $\Sigma h_i/20$; $\mu(2)$ = expected central moment, under $H_0$, see (3); $L$ = upper 95% confidence limit, see (23); * indicates $s^2_K(1)$ exceeds $L$, (*) indicates $s^2_K(2)$ exceeds $L$.

$s^2_K$ = observed variance of the number of heterozygous comparisons; (1) raw data, (2) raw data after correction (formula 24), (3) subset of complete data.

$X(2)$, $X(3)$, $X(4)$ = measures of multilocus structure, see (formula 19).

† $\mu(3)$ negative.

‡ $m(3)$, $\mu(3)$ negative.

phism. The average for these populations was 0.80. Some populations with high values (*e.g.*, 24, Akhziv; 11, Damon) were from disturbed or temporary habitats. These samples comprised relatively few multilocus genotypes that were differentiated from one another at several loci. The Bor Mashash (17) sample was dominated by one multilocus genotype, with a few others in low frequency. This suggested a highly localized population, confined to the more favorable microhabitats in this site from the Negev desert. The populations on the coastal plain (24–28) showed a cline of decreasing values of $X(2)$, but generally there was no overall correlation with latitude.

Sample sizes were probably inadequate for individual estimates of $X(3)$ and $X(4)$. Overall, there was no significant trend to positive or negative values of $X(3)$. However, the values of $X(4)$ were mostly positive, and correlated with $X(2)$.

Table 2 summarizes estimates of Spearman's rank correlation between the variables $s^2_K$ or $X(2)$ with the number of polymorphic loci ($m'$), the mean genetic diversity ($\bar{h}$), the average genetic distance or geographic distance between a population and all other 28 populations (NEVO *et al.* 1979) and the mean annual rainfall. Clearly, the variance in heterozygosity is correlated with the two single-locus measures of genetic variation whereas the standardized ratio $X(2)$ is not. This result supports the standardization procedure (19).

This standardization yielded a novel and interesting result, namely, a marked association between mean genetic distance and intensity of multilocus association. In part, this arises because genetic diversity ($\bar{h}$) and genetic distance are associated ($r_s = -0.56$). Yet, the correlations suggest that populations with low diversity tend to be genetically more distinct from all other populations and to exhibit more intense multilocus associations. Since geographic remoteness was apparently unrelated to genetic distance ($r_s = -0.08$), this syndrome of characters was not primarily due to isolation by distance, but might be typical for colonial or marginal or small populations.

*Multilocus analysis of four esterase loci in wild and cultivated barley:* Perhaps the most extensive data on multilocus associations are those of CLEGG, ALLARD and KAHLER (1972) for composite crosses II and V in barley (*H. vulgare*). They used a hierarchical chi-square analysis to demonstrate the increase in multilocus structure with generations. However, the chi-square values are also functions of sample size, which fortuitously increased with increasing generations. SMOUSE (1974) analyzed the gametic frequencies of one of the generations (CCV, $F_{26}$), using a log linear model. In his analysis, most of the multiple locus disequilibrium was accounted for by two-locus effects.

It seemed desirable to compare the intensity of association found in these cultivated barley populations with the levels observed in wild barley. To do this,

TABLE 2

*Rank correlation* ($r_s$) *between two measures of association and genetic and environmental variables*

| | Variance in heterozygosity ($s^2_K$) | Association intensity $X(2)$ |
|---|---|---|
| Loci polymorphic ($m'$) | 0.47* | —0.15 |
| Genetic diversity ($\bar{h}$) | 0.50** | —0.18 |
| Mean genetic distance† | 0.15 | 0.55** |
| Mean geographic distance† | 0.18 | 0.15 |
| Mean annual rainfall | —0.04 | 0.10 |

*, ** Denote statistically significant correlation at the 0.05 and 0.01 levels, respectively.
† From each population to all other populations (see NEVO *et al.* 1979).

we recomputed the statistics of Table 1 using only the data for the esterase loci. For the $s^2_K$, we used the adjusted value, method (2) above. Table 3 shows the average value of the 26 polymorphic populations, and individual values for the population that showed the maximum $X(2)$ for esterases, Bor Mashash. The four-locus gamete frequencies for the composites computed from CLEGG, ALLARD and KAHLER (1972), were also analyzed. It is important to note that CLEGG, ALLARD and KAHLER had grouped the alleles into the most common and the remainder, and in so doing used only diallelic scores at each locus. Higher values for $h$, $\mu(2)$ and $s^2_K$ would result if the polyallelic scores were used, but this would not necessarily increase the estimate of $X(2)$.

The intensity values, $X(2)$, do confirm CLEGG, ALLARD and KAHLER's claim that the degree of multilocus structure has increased during the evolution of the composites. The last sample of both crosses show marked increases of values of $X(2)$ over earlier generations. In CCII, this is accompanied by a decline in diversity ($\bar{h}$). Because of the large sample sizes, all the parameters are estimated very accurately. The values for $X(3)$, related to the skewness of the basic distribution, show no fixed pattern. Apart from $F_{18}$ of CCII, they are apparently dominated by two-locus effects.

The values of $X(2)$ and $X(3)$ for esterase loci in the Israeli wild barley populations indicate a higher degree of multilocus association than in the composite crosses. BROWN, NEVO and ZOHARY (1977) presented a preliminary analysis of some of these populations based on information statistics. It was argued that such associations are the corollary of co-adaptation, which, if present, strengthens the case for the use of genetic conservation as opposed to mutagenesis as a source of variation for future breeding. It is recognized that historic effects or other scenarios of indirect selection could also account for multilocus associations.

TABLE 3

*Multilocus association among four esterase loci in two composite crosses of barley;* H. vulgare *(data of* CLEGG, ALLARD and KAHLER *1972) compared with* Hordeum spontaneum *populations*

| *Hordeum vulgare* Population | Generation | $n$ | $\bar{h}$ | $\mu(2)$ | $L$† | $s^2_K$ | $X(2)$ | $X(3)$ |
|---|---|---|---|---|---|---|---|---|
| CC II | 7 | 1,044 | 0.42 | 0.93 | 1.02** | 1.20 | 0.29 | 1.8 |
| | 18 | 2,087 | 0.41 | 0.92 | 0.98** | 1.07 | 0.17 | —2.4 |
| | 41 | 2,868 | 0.22 | 0.68 | 0.72** | 1.66 | 1.44 | 7.0 |
| CC V | 5 | 1,452 | 0.40 | 0.87 | 0.93 | 0.91 | 0.04 | 0.3 |
| | 17 | 2,443 | 0.40 | 0.88 | 0.92* | 0.93 | 0.05 | 0.2 |
| | 26 | 3,049 | 0.44 | 0.98 | 1.03** | 1.19 | 0.22 | 0.5 |
| *Hordeum spontaneum* | $m'$ | | | | | | | |
| Average Israeli population | 3 | 43 | 0.30 | 0.61 | 0.92** | 0.96 | 0.54 | 2.8 |
| Highest (17) | 4 | 52 | 0.16 | 0.49 | 0.76** | 1.27 | 1.58 | 7.6 |

† $L$ is the confidence limit: if it is shown with **, it is the 99% limit; otherwise, it is the 95% limit.

## CONCLUSIONS

From an initial suggestion of SVED (1968), we developed a set of indices of multilocus organization, and exemplified them with data of wild and cultivated barley. The measures are based on the observed distribution of the number of heterozygous loci in two *randomly chosen gametes*, $f(K)$. Note that unless one is dealing with ideal, panmictic populations, this distribution does *not* correspond with that of the number of heterozygous loci in a sample of diploid zygotes. The most useful statistic from the basic empirical distribution is the variance in number of heterozygous loci ($s^2_K$). The expected value for this variance, given the single-locus genetic diversities and assuming that loci are independent, [$\mu(2)$] and its sampling variance are readily computed. A test of the null hypothesis of "no inflation of the variance due to association" is possible.

The measures have a number of distinct advantage in comparison with using the set of $D$ parameters, log linear models oi information statistics. First, the underlying distribution, $f(K)$, is relatively simple to calculate, and the measures directly follow from this as observed moments. Second, they can be used with relatively small samples of organisms (in the order of 30 here), especially if a large number of loci are scored. Third, they effectively summarize multilocus association to a few values for comparative studies among populations or species. Fourth, this mode of information reduction emphasizes one important aspect of multilocus organization, namely the occurrence of multilocus heterozygosity. This aspect has considerable biological appeal when compared with other summarizing methods, such as $\sum_{ij}\sum \mathrm{D}^2_{ij}$ or $\sum_{ij}\sum \mathrm{D}^2_{ij}/p_ip_j$. Fifth, the variance is a function of the sum of squares of all possible two-locus gametic frequencies (see formula 16). Thus, this measure of "evenness" is an inverse function of two-locus gametic frequencies in inverse analogy with the single-locus SIMPSON diversity measure.

Of course, there are several drawbacks to these measures that should be kept in mind. First, such an extreme summary of data on several loci represents a severe loss of information. It ignores the behavior of particular allelic combinations. The study of such combinations requires methods based on $D$ statistics, as recently reviewed by WEIR (1979). Yet, the study of weak associations between particular alleles frequently requires sample sizes beyond those normally considered realistic for electrophoretic studies (BROWN 1975). Second, the measures assume each locus is of equal interest. While this assumption is commonly made in selectively neutral models or models of general heterozygous advantage, it is not realistic biologically. For example, associations involving one particular locus, such as an alcohol dehydrogenase locus, could be of more biological importance than associations among a closely linked group of markers, such as some barley esterases. Third, formula (12) shows that unless allele frequencies are all equal, there can exist configurations of disequilibrium for which $\sigma^2_K$ is *lower* than its value when the loci are independent. In such cases, summing over pairs of loci as in formula (15) may amount to combining positive and negative contributions. Therefore, the null hypothesis that an observed value ($s^2_K$)equals the expected value of $\sigma_K^2$ under independence includes conditions when some multilocus organization may indeed be present. Fourth, the measure of intensity of multilocus structure cannot readily be made entirely independent of single-locus diversity or the number of loci scored. This property should be borne in mind when comparing, from different species, estimates that are based on radically different single-locus data bases.

A. H. D. BROWN is grateful for a travel grant that permitted completion of this work. We thank L. ALBRECHT for excellent computational assistance, and M. T. CLEGG, J. SVED, B. S. WEIR and a reviewer for important suggestions.

## LITERATURE CITED

AVERY, P. J. and W. G. HILL, 1979 Variance in quantitative traits due to linked dominant genes and variance in heterozygosity in small populations. Genetics **91**: 817–844.

BROWN, A. H. D., 1975 Sample sizes required to detect linkage disequilibrium between two and three loci. Theoret. Pop. Biol. **8**: 184–201.

BROWN, A., E. NEVO and D. ZOHARY, 1977 Association of alleles at esterase loci in wild barley *Hordeum spontaneum*. Nature **268**: 430–431.

BROWN, A. H. D., E. NEVO, D. ZOHARY and O. DAGAN, 1978 Genetic variation in natural populations of wild barley (*Hordeum spontaneum*). Genetica **49**: 97–108.

BROWN, A. H. D., D. ZOHARY and E. NEVO, 1978 Outcrossing rates and heterozygosity in natural populations of *Hordeum spontaneum* Koch in Israel. Heredity **41**: 49–62.

CLEGG, M. T., R. W. ALLARD and A. L. KAHLER, 1972 Is the gene the unit of selection? Evidence from two experimental plant populations. Proc. Natl. Acad. Sci. U.S. **69**: 2474–2478.

CLEGG, M. T., J. F. KIDWELL, M. G. KIDWELL and N. J. DANIEL, 1976 Dynamics of correlated genetic systems. I. Selection in the region of the Glued locus of *Drosophila melanogaster*. Genetics **83**: 793–810.

FRANKLIN, I., 1977 The distribution of the proportion of the genome which is homozygous by descent in inbred individuals. Theoret. Pop. Biol. **11**: 60–80.

HEDRICK, P., S. JAIN and L. HOLDEN, 1978 Multilocus systems in evolution. Evolutionary Biol. **11**: 101–184.

HILL, W. G., 1974 Estimation of linkage disequilibrium in randomly mating populations. Heredity **33**: 229–239. ——, 1975 Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. Theoret. Pop. Biol. **8**: 117–126.

KARLIN, S., 1975 General two-locus selection models. Some objectives, results and interpretations. Theoret. Pop. Biol. **7**: 364–398.

KENDALL, M. G. and A. STUART, 1969 *The Advanced Theory of Statistics*, Vol. 1. *Distribution Theory*. Third Edition. Charles Griffin, London.

NEVO, E., D. ZOHARY, A. H. D. BROWN and M. HABER, 1979 Genetic diversity and environmental associations of wild barley, *Hordeum spontaneum* in Israel. Evolution **33**: 815–833.

SIMPSON, E. H., 1959 Measurement of diversity. Nature **163**: 688.

SMOUSE, P. E., 1974 Likelihood analysis of recombinational disequilibrium in multiple locus gametic frequencies. Genetics **76**: 557–565.

SVED, J. A., 1968 The stability of linked systems of loci with a small population size. Genetics **59**: 543–563.

WEIR, B. S., 1979 Inferences about linkage disequilibrium. Biometrics **35**: 235–254.

Corresponding editor: B. WEIR