A. Charcosset · L. Essioux

The effect of population structure on the relationship between heterosis and heterozygosity at marker loci

Received: 8 October 1993 / Accepted: 2 February 1994

Abstract The relationship between heterozygosity at neutral marker loci and heterosis of F_1 hybrids is investigated using a theoretical model. Results emphasize that linkage disequilibrium between the markers and the loci implicated in heterosis [quantitative trait loci (QTLs) that exhibit dominance effects] is a necessary condition to finding a correlation (ρ_{mh}) between heterozygosity at marker loci and the heterosis. The effect of population structure, in which the parental inbred lines of the hybrids belong to different heterotic groups, is considered. ρ_{mh} is investigated for: (1) hybrids between lines that belong to the same heterotic group (within-group hybrids); (2) hybrids between lines that belong to different groups (between-group hybrids); and (3) all hybrids, both within and between-groups. Within a group, significant values of (ρ_{mh}) may arise because of linkage disequilibrium generated by drift. At the between-group level, no correlation is expected since linkage disequilibrium should differ randomly from one group to the other, which is consistent with recent experimental results. Possible ways to achieve prediction of the heterosis in this situation are discussed. When all hybrids are considered simultaneously, divergence of allelic frequencies among groups for the markers and the QTLs produces a correlation between heterosis and heterozygosity at marker loci. This correlation increases with the number of markers that are considered.

Key words Distance · Markers · Heterozygosity · Heterosis · Specific combining ability

Communicated by A. R. Hallauer

A. Charcosset (✉) · L. Essioux
INRA-UPS-INAPG, Station de Génétique Végétale, Ferme du Moulon 91190 Gif/Yvette, France

Introduction

When introducing the basic concepts of hybrid breeding, Shull (1908) noted that finding a suitable method to predict hybrid performance before field evaluation would considerably increase the efficiency of maize breeding programs. Experimental studies (see Moll et al. 1965) illustrated that midparent heterosis could be related to genetic divergence, a relationship that was supported by quantitative genetics theory (Falconer 1981). On the basis of these results, the efficiency of several distance indicators for predicting either heterosis or hybrid value was evaluated.

Special interest was devoted to genetic markers, first using isozyme data, and then molecular markers. The results obtained using isozyme data were reviewed by Stuber (1989). Distances computed from isozyme data were in some cases significantly correlated to heterosis, but the correlations were generally too low for the distances to be of practical predictive value. Insufficient genome coverage, due to the low number of marker loci, was a possible explanation for these results. Alternative molecular marker techniques, such as restriction fragment length polymorphism (RFLP) have been developed over the last decade, (Beckmann and Soller 1983; Burr et al. 1983), and these provide a much higher number of genetic markers (up to 1000 in some major crops). The potential of this technique for prediction of either heterosis or hybrid value has been tested in several studies in maize since 1988 (Lee et al. 1989; Godshalk et al. 1990; Melchinger et al. 1990 a, b; Smith et al. 1990; Dudley et al. 1991; Melchinger et al. 1992; Boppenmeier et al. 1992; Charcosset 1992), with the results appearing to be highly dependent on the germ plasm involved in the study. A general tendency seems that the correlation between distances (computed using RFLP data) and hybrid value increases with the introduction of crosses between related lines in the germ plasm under study. This tendency was also pointed out for isozymes by Frei et al. (1986). This effect of relatedness is consistent with results obtained by Stuber et al. (1992) that showed a strong correspondence between heterozygosity at marker loci and yield when backcross families were considered.

The relationship between heterosis and heterozygosity at marker loci is also affected by the origin of the tested germ plam; i.e., the heterotic groups (e.g. Lancaster and Reid Yellow Dent for maize) to which the lines belong. Melchinger et al. (1992) concluded that: (1) RFLP distances should be useful for predicting hybrid value when considering diallels that involve lines that belong to a same heterotic group, or lines that belong to several heterotic groups, but (2) that RFLP distances would not be useful for predicting the value of crosses between lines that belong to different heterotic groups. This conclusion was confirmed by Boppenmeier et al. (1992).

At the genetic level, several factors affect the relationship between marker distance and hybrid performance. Since markers are generally assumed to be neutral, linkage disequilibrium between markers and the loci involved in heterosis is a necessary condition (Charcosset et al. 1991). Other factors are related to: (1) the degree of dominance at the loci [quantitative trait loci (QTL)] involved in the variation of the quantitative trait of interest and (2) epistatic effects (Melchinger et al. 1990b). Apropos to previous results, elucidating the effect of population structure (i.e., the partition of inbred lines into groups on the basis of origins) on the relationship between distance and hybrid performance is essential to the stipulation of conditions under which marker-based prediction will be effective. Bernardo (1992) provided theoretical arguments on that issue using simulated data. The aim of this paper is to develop an analytical approach by extending the study of Charcosset et al. (1991). In the present study we will consider the following three cases: (1) diallels that involve lines belonging to the same genetic group, (2) factorial designs in which lines of a first group are crossed to lines of a second group, and (3) diallels involving lines belonging to different groups.

Basis of the model

Quantitative traits

We shall assume a biallelic model to describe the phenotypic value of homozygous inbred lines and their hybrids, following the notations used in a previous paper (Charcosset et al. 1991). Using the notation of Hayman (1954), the genotype of individual i at locus l (with alleles l_1 and l_2) is represented by the variable θ_l^i, which takes the value $+1, 0, -1$ for genotypes l_1l_1, l_1l_2, and l_2l_2, respectively. The single-locus model for the phenotypic value of individual i is written as $Y_i = c_l + a_l\theta_l^i + d_l(1-(\theta_l^i)^2)$, where c_l is the average value of homozygotes l_1l_1 and l_2l_2, a_l is half the difference between the homozygotes l_1l_1 and l_2l_2 phenotypes, and d_l is the difference between the heterozygote phenotype (l_1l_2) and the average of homozygotes, i.e., the dominance effect. If the trait is controlled by nl loci acting independently (no epistasis), the phenotype of individual i (Y_i) is (with $C = \sum_l^{nl} c_l$) written as

$$Y_i = C + \sum_{l=1}^{nl} a_l(\theta_l^i) + d_l(1-(\theta_l^i)^2). \tag{1}$$

Note that, when i is an inbred line homozygote for each QTL, its value (per se) reduces to: $Y_{ii} = C + \sum_{l=1}^{nl} a_l\theta_l^i$. The value of the F_1 hybrid between line i and line j is

$$Y_{ij} = C + \sum_{l=1}^{nl} a_l\frac{(\theta_l^i + \theta_l^j)}{2} + d_l\frac{(1-\theta_l^i\theta_l^j)}{2}. \tag{2}$$

Let He_{ij} be the difference between the hybrid value and the mean of the values of the parents (i.e., the heterosis), then

$$He_{ij} = \sum_{l=1}^{nl} d_l\frac{(1-\theta_l^i\theta_l^j)}{2}. \tag{3}$$

Assuming that d_l equals d whatever l ($\forall_l d_l = d$), heterosis is proportional to the average heterozygosity at QTL loci (Falconer 1981). Variation in d_l values will affect this relationship, as will be discussed further. Epistatic effects were not included in the model used for this study and deserve a specific analysis since they modify the relationship between heterosis and the average heterozygosity at the QTLs. Crow and Kimura (1970, p 81) illustrated that, in the presence of epistatic effect, the relationship is no longer linear but gets curvilinear, with a concavity that depends on the type and magnitude of the epistatic effects involved.

Since the average heterozygosity of a given hybrid can be estimated via marker loci analysis of parental inbred lines, under the hypothesis that this estimate of heterozygosity is proportional to the number of heterozygous QTL, average heterozygosity has been considered to be a predictor of heterosis.

Heterozygosity at marker loci

When np marker loci are available, distances between inbred lines i and j can be computed using well-known formulas, such as MRD^2 (Rogers 1972). If inbred lines are homozygous for all loci (which will be assumed in this study), this distance is an estimate of the average heterozygosity of the hybrid between lines i and j and will be designated MD_{ij} (for marker distance). For a given marker locus (p), variable θ_p takes values -1, $+1$ for genotypes p_1p_1 and p_2p_2, respectively. Following that notation,

$$MD_{ij} = \frac{\sum_{p=1}^{np}(1-\theta_p^i\theta_p^j)}{2np}. \tag{4}$$

Note that this model is adapted to describe situations in which more than two alleles are detected at marker loci (which is a general situation for RFLPs in maize). In this case, variable $\theta^i_{p_a}$ is defined for allele a of locus p, taking the value 1 if the inbred i carries allele p_a, and -1 if not. If na is the total number of alleles found over the np marker loci, $MD_{ij}=\sum_{pa=1}^{na}(1-\theta^i_{pa}\theta^j_{pa})/4np$.

Genetic parameters of the reference population

We will assume that the homozygous inbred lines belong to a reference population (i.e., a set of inbreds of infinite size). We will define w_l as the mean of θ^i_l in this population. The frequency of the allele l_1 in the population is: $f_{l_1}=(1+w_l)/2$. Genetic diversity (Nei 1973) at locus l (H_l) is proportional to the variance of θ^i_l: $H_l=var(\theta^i_l)/2=(1-w_l^2)/2$. The linkage disequilibrium between alleles l_1 and k_1 at loci l and k is proportional to the covariance between variables θ^i_l and θ^i_k: $D_{lk}=cov(\theta^i_l;\theta^i_k)/4$. As was discussed by several authors (see Roughgarden 1979, p 113) the name linkage disequilibrium may be misleading in that sense that D_{lk} is not necessarily due to linkage, as will be discussed further. However, linkage disequilibrium is the name that is the most commonly used for the statistical association between two alleles, so it will be used to refer to D_{lk} in the following text.

To investigate the effect of population structure, we will consider that the population (or metapopulation, following the terminology of population genetics) is divided into several subpopulations, as is the case when the population is divided into several (ng) heterotic groups. These subpopulations will be designated as $(G1, G2 \ldots Gng)$. Previously defined population parameters can be defined for each subpopulation (g) as $w^g_l, f^g_{l_1}, H^g_l$ and D^g_{lk}. The relative size of group g is defined as $f_g(\sum_{g=1}^{ng} f_g=1)$.

Relationship between heterosis and heterozygosity at marker loci in diallel mating designs

Variation within a diallel is partitioned (Griffing 1956) as

$$Y_{ij}=\mu+GCA_i+GCA_j+SCA_{ij}, \tag{5}$$

where Y_{ij} is the value of the hybrid between inbreds i and j, μ is the mean over the hybrids, GCA_i and GCA_j are the general combining ability of inbreds i and j, respectively; SCA_{ij} is the specific combining ability of the hybrid ij, i.e., the specific heterosis (Gardner and Eberhart 1966).

When Griffing's (1956) method I is used (this method leads to simplified calculations and doesn't affect the results, since the size of the population is supposed to be infinite), the specific combining ability between lines i and j is

$$SCA_{ij}=\frac{1}{2}\sum_{l=1}^{nl} d_l(\theta^i_l-w_l)(w_l-\theta^j_l). \tag{6}$$

Application of model (5) to marker distance (MD) (Melchinger et al. 1990b) provides a quantity expressing the specific marker distance (SMD) between inbreds i and j:

$$SMD_{ij}=\frac{1}{2np}\sum_{p=1}^{np}(\theta^i_p-w_p)(w_p-\theta^j_p). \tag{7}$$

In the present study, we will analyze the relationship between He_{ij} and MD_{ij} through the investigation of the relationship between SCA_{ij} and SMD_{ij}. This is justified because SCA is the most important component of heterosis concerning the relationship with marker information, since other components of heterosis [or general heterosis according to Gardner and Eberhart (1966)] can be predicted using top-cross designs (Sprague and Tatum 1942). Given that SCA is the relevant component of heterosis it seems natural to regard SMD as the relevant component of marker distance. The calculation of the correlation between SCA_{ij} and SMD_{ij} $(\rho(SCA_{ij}; SMD_{ij}))$ requires three components: $cov(SCA_{ij}; SMD_{ij})$, $var(SCA_{ij})$, and $var(SMD_{ij})$. The variance of SCA_{ij} (see Appendix) is

$$var(SCA_{ij})=\sum_{k=1}^{nl} d_k^2 H_k^2+4\sum_{k=1}^{nl}\sum_{l=1,l\neq k}^{nl} d_k d_l D_{lk}^2. \tag{8}$$

Equation 8 shows that the variance of specific combining ability depends on the diversity at the QTLs and linkage disequilibrium between the QTLs. If d_l is greater than 0 at all loci (l), linkage disequilibrium will tend to increase the variance of SCA. Similarly,

$$var(SMD_{ij})=\frac{1}{np^2}\sum_{p=1}^{np} H_p^2+\frac{4}{np^2}\sum_{p=1}^{np}\sum_{q=1,q\neq p}^{np} D^2_{pq}, \tag{9}$$

and

$$cov(SCA_{ij}; SMD_{ij})=\frac{4}{np}\sum_{k=1}^{nl}\sum_{p=1}^{np} d_k D^2_{kp}. \tag{10}$$

Equation 10 illustrates that linkage disequilibrium between neutral markers and QTLs is a necessary condition for correlation between heterosis and heterozygosity at marker loci. Several aspects of this problem were discussed in a previous paper (Charcosset et al. 1991), under the assumption that inbred lines were derived from a population that had undergone random mating for a given number of generations after foundation. The results indicated that the relationship was expected to be at a maximum in the first generations, which is consistent with the high values of the correlation reported by Stuber et al. (1992) for backcross families derived from a cross between two inbred lines.

If (1) dominance is constant across all QTLs, (2) diversity is the same for all the loci (QTLs and markers), (3) there is no linkage disequilibrium between the QTLs, (4) there is no linkage disequilibrium between the marker loci, and (5) linkage disequilibrium between a QTL and a given marker is either maximum ($D^2 = H^2/4$) or zero,

$$\rho^2(SCA_{ij}; SMD_{ij}) = (\%MQTL)(\%QAM), \qquad (11)$$

where $\%MQTL$ is the proportion of the QTLs marked by a marker (i.e., associated with a marker) and $\%\,QAM$ is the proportion of the markers associated with a QTL. This clearly illustrates that heterosis estimates based on marker heterozygosity will be inflated by heterozygous markers unassociated with QTLs [dispersed markers, Bernardo (1992)]. On the other hand, if heterotic QTLs are unassociated with markers, heterosis will be underestimated by estimates based on marker heterozygosity. Thus, dispersed markers and unmarked QTLs play symmetric roles.

The correlation between SCA and SMD is also affected by the variation of the dominance effect across the QTLs. If one assumes that (1) diversity is the same for all the QTLs, (2) there is no linkage disequilibrium between the QTLs, (3) each QTL is associated with a single marker with maximum linkage disequilibrium ($D^2 = H^2/4$), and (4) there are no dispresed markers, the correlation is inversely related to the variation of d_l across the QTLs: $\rho^2(SCA_{ij}; SMD_{ij}) = 1/(1 + \frac{var_d}{mean_d^2})$, where $mean_d$ and var_d are, respectively, the mean and the variance of d_l across the QTLs.

Relationship between heterosis and heterozygosity at marker loci in factorial mating designs (between-groups' hybrids)

When lines of a given group ($G1$) are crossed to lines of another group ($G2$), in a factorial design (also called design II by Hallauer and Miranda 1988), variation can be partitioned via the model

$$Y_{ij}^{12} = \mu + GCA_i^1 + GCA_j^2 + SCA_{ij}^{12}, \qquad (12)$$

where Y_{ij}^{12} is the value of the hybrid between inbred i in group $G1$ and inbred j in group $G2$, μ is the mean over the hybrids, GCA_i^1 is the general combining ability of inbred i in group $G1$ with the inbreds of group $G2$, GCA_j^2 is the general combining ability of inbred j in group $G2$ with the inbreds of group $G1$; SCA_{ij}^{12} is the specific combining ability of the hybrid ij.

As in the case of the diallel, model 12 can be applied both to the quantitative trait of interest and the marker distance between lines i and j. The values of SCA_{ij}^{12} and SMD_{ij}^{12} are

$$SCA_{ij}^{12} = \frac{1}{2}\sum_{l=1}^{nl} d_l(\theta_l^i - w_l^1)(w_l^2 - \theta_l^j), \qquad (13)$$

and

$$SMD_{ij}^{12} = \frac{1}{2}\sum_{p=1}^{np} (\theta_p^i - w_p^1)(w_p^2 - \theta_p^j). \qquad (14)$$

Similarly, the components of the correlation between SCA_{ij}^{12} and SMD_{ij}^{12} are

$$var(SCA_{ij}^{12}) = \sum_{k=1}^{nl} d_k^2 H_k^{g1} H_k^{g2} + 4\sum_{k=1}^{nl}\sum_{l=1, l\neq k}^{nl} d_k d_l D_{lk}^{g1} D_{lk}^{g2} \qquad (15)$$

$$var(SMD_{ij}^{12}) = \sum_{p=1}^{np} H_p^{g1} H_p^{g2} + 4\sum_{p=1}^{np}\sum_{q=1, q\neq p}^{np} D_{pq}^{g1} D_{pq}^{g2} \qquad (16)$$

$$cov(SCA_{ij}^{12}; SMD_{ij}^{12}) = 4\sum_{k=1}^{nl}\sum_{p=1}^{np} d_k D_{kp}^{g1} D_{kp}^{g2} \qquad (17)$$

Formula 17 illustrates a major difference between the diallel and factorial mating designs. In the diallel design, the covariance between SCA and SMD involves the square of linkage disequilibria between marker loci and QTLs so that any disequilibrium contributes positively to the correlation between SCA and SMD. For factorial designs, the covariance is the product of the disequilibria observed in the two groups. Thus, disequilibria of opposite sign in the two populations will contribute in a negative way to the correlation, as was pointed out by Melchinger (1991, unpublished data reported by Boppenmeier et al. 1992).

The effect of germ plasm structure on the relationship between heterosis and heterozygosity at marker loci in diallel designs

Diallels often involve lines that represent different heterotic groups. Thus, these designs can be considered as a mixture of within-group diallels and factorial designs (between-groups' hybrids). We will consider a population of inbred lines subdivided into ng groups. Following an approach similar to that of Ohta (1982), linkage disequilibrium can be partitioned as

$$D_{lk} = \sum_{g=1}^{ng} f_g D_{lk}^g + \frac{1}{4}\sum_{g=1}^{ng} f_g(w_l^g - w_l)(w_k^g - w_k). \qquad (18)$$

Equation 18 illustrates that linkage disequilibrium is the result of two kinds of effects: (1) the pooled within-population disequilibria (over the groups) and (2) the divergence of the groups for allelic frequencies as $(w_l^g - w_l)$ is equal to twice the difference between the frequency of allele l_1 in group g and its frequency in the metapopulation.

To illustrate the effect of allelic divergence of the groups on the correlation between SCA and SMD, we will consider the case of no linkage disequilibrium within each population (i.e., each group is at linkage equilibrium) and only two populations equal in size. Then,

$$D_{lk} = \frac{1}{16}(w_l^{g1} - w_l^{g2})(w_k^{g1} - w_k^{g2})$$

$$= \frac{1}{4}(f_{l_1}^{g1} - f_{l_1}^{g2})(f_{k_1}^{g1} - f_{k_1}^{g2}). \qquad (19)$$

Following Eq. 10 the covariance between *SCA* and *SMD* is

$$cov(SCA_{ij}; SMD_{ij}) = \frac{1}{4np} \sum_{l=1}^{nl} \sum_{p=1}^{np} d_l (f_{l_1}^{g1} - f_{l_1}^{g2})^2 (f_{p_1}^{g1} - f_{p_1}^{g2})^2. \tag{20}$$

Assuming that all QTLs have the same dominance effect ($\forall_l d_l = d$), Eq. 20 is expressible as a function of the divergence of the allelic frequencies between the populations at the QTLs ($\Delta_q = 1/nl \sum_{l=1}^{nl} (f_{l_1}^{g1} - f_{l_1}^{g2})^2$) and the marker loci ($\Delta_m = 1/np \sum_{p=1}^{np} (f_{p_1}^{g1} - f_{p_1}^{g2})^2$),

$$Cov(SCA_{ij}; SMD_{ij}) = \frac{nld}{4} \Delta_q \Delta_m. \tag{21}$$

Equation 21 illustrates that the diversity in allelic frequencies between the populations at the markers and the QTLs will tend to generate linkage disequilibria between marker loci and QTLs, even if those are not physically linked, thereby generating a correlation between *SCA* and *SMD*.

The diversity at locus *l* can be written: $H_l = \frac{1}{2}((w_l^{g1} - w_l^{g2})/2)^2 + \frac{1}{2}(1 - ((w_l^{g1})^2 + (w_l^{g2})^2)/2) = \frac{1}{2}\delta_l + \bar{H}_l$, where δ_l is the square of the differences in the allelic frequencies of groups *G*1 and *G*2 at locus *l*, $\bar{H}_l$ is the average diversity of groups *G*1 and *G*2. Consequently, using Eqs. 8 and 19, *var*(*SCA*) becomes

$$var(SCA_{ij}) = d^2 \frac{nl^2}{4} \Delta_q^2 + d^2 \sum_{l=1}^{nl} \bar{H}_l(\bar{H}_l + \delta_l). \tag{22}$$

Similarly,

$$var(SMD_{ij}) = \frac{1}{4} \Delta_m^2 + \frac{1}{np} \sum_{p=1}^{np} \bar{H}_p(\bar{H}_p + \delta_p), \tag{23}$$

and the correlation between *SCA* and *SMD* becomes $\rho(SCA_{ij}; SMD_{ij}) = \Gamma_q \Gamma_m$, with

$$\Gamma_q = \frac{\Delta_q}{\sqrt{\Delta_q^2 + \frac{4}{nl^2} \sum_{l=1}^{nl} \bar{H}_l(\bar{H}_l + \delta_l)}}$$

and

$$\Gamma_m = \frac{\Delta_m}{\sqrt{\Delta_m^2 + \frac{4}{np^2} \sum_{p=1}^{np} \bar{H}_p(\bar{H}_p + \delta_p)}}.$$

The fraction of the variance of specific combining ability that is accounted for by the heterozygosity at marker loci is

$$\rho^2(SCA_{ij}; SMD_{ij}) = \Gamma_q^2 \Gamma_m^2. \tag{24}$$

Since $\bar{H}_l(\bar{H}_l + \delta_l) = H_l^2 - \frac{1}{4}\delta_l^2$ Eq. (24) illustrates that the magnitude of the relationship between *SCA* and *SMD* depends on the degree to which diversity in both markers and QTLs is distributed between groups.

Another way to derive the parameters Γ_q and Γ_m is to consider a variable G_{ij} such that $G_{ij} = 0$ if inbreds *i* and *j* belong to the same group and $G_{ij} = 1$ if they belong to different groups. It can be demonstrated that $\Gamma_q = \rho(SCA_{ij}; G_{ij})$. Thus, Γ_q is indicative of the accuracy of the prediction of *SCA* from the knowledge of the groups to which the inbreds belong. In a similar way, $\Gamma_m = \rho(G_{ij}; SMD_{ij})$. Thus, Γ_m is indicative of the accuracy marker distance to determine if two inbreds belong or do not belong to the same group.

To illustrate this result, we will consider that diversity and allelic divergence are the same for all the QTLs (i.e., $\forall_l H_l = H_q, \delta_l = \delta_q$) and for all the markers (i.e., $\forall_p H_p = H_m, \delta_p = \delta_m$). In this situation $\Gamma_q^2 = (\delta_q^2/(\delta_q^2(1 - 1/nl) + (4/nl)H_q^2))$ and $\Gamma_m^2 = (\delta_m^2/(\delta_m^2(1 - 1/np) + (4/np)H_m^2))$.

Table 1 shows the values of $\Gamma^2 = \delta^2/(\delta^2(1 - 1/n) + (4/n)H^2)$ for various values of the allelic frequencies in groups *G*1 (f_1) and *G*2 (f_2), under the hypothesis that all loci exhibit the same diversity and divergence. The value of Γ increases with the divergence of the populations ($|f_1 - f_2|$). For a given divergence, Γ increases as the average within-group diversity decreases and is maximum at fixation for either population. For given values of the diversity and divergence, Γ increases with the number of loci that considered. Table 1 allows computation of the fraction of the variance of specific combining ability that is accounted for by the heterozygosity at marker loci under various hypothesis concerning the number of loci, the diversity and the divergence of the groups for the markers, and the QTLs. It illustrates that the divergence of the groups can generate high correlation values, especially if the number of marker loci that are considered is important.

Discussion and conclusion

Supposing that markers such as RFLPs and isozymes are neutral [i.e., have no direct effect on the trait(s) of interest], then they must be in linkage disequilibrium with QTLs to have a predictive value. This is illustrated by formula 11 using simplifying assumptions: if 50% of the markers are dispersed and 50% of the QTLs are unmarked, the fraction of the variance of *SCA* that is accounted for by heterozygosity at marker loci will not exceed 25%. In addition to linkage disequilibrium parameters, the correlation between heterosis and heterozygosity at marker loci is affected by the genetics of the trait of interest. Variation in the value of the dominance effect across the QTLs and epistatic effects contribute to diminish the correlation and deserve further investigations. However, linkage disequilibrium is the key parameter by which to investigate the effect of population structure on the effectiveness of prediction. Concerning linkage disequilibria between markers and QTLs for structured populations, two parameters are paramount: (1) linkage disequilibrium within the heterotic groups of interest and (2) divergence of the groups in allelic frequencies.

Table 1 Values of parameter Γ^2 for given allelic frequencies in groups $G1$ ($f1$) and $G2$ ($f2$), and two number of loci ($n = 10$ and $n = 100$). The fraction of the variance of specific combining ability (*SCA*) which is accounted for by the heterozygosity at marker loci (*SMD*) is: $\rho^2(SCA; SMD) = \Gamma_q^2\Gamma_m^2$, where parameters Γ_q^2 and Γ_m^2 are defined for the QTLs and the marker loci, respectively. (*nd* Γ^2 undefined due to the absence of diversity)

$n = 10$	f2										
f1	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.0	nd	0.027	0.111	0.243	0.400	0.556	0.692	0.803	0.889	0.953	1.000
0.1	0.027	0.000	0.004	0.038	0.127	0.274	0.449	0.621	0.764	0.874	0.953
0.2	0.111	0.004	0.000	0.002	0.022	0.090	0.222	0.405	0.598	0.764	0.889
0.3	0.243	0.038	0.002	0.000	0.001	0.017	0.077	0.208	0.405	0.621	0.803
0.4	0.400	0.127	0.022	0.001	0.000	0.001	0.016	0.077	0.222	0.449	0.692
0.5	0.556	0.274	0.090	0.017	0.001	0.000	0.001	0.017	0.090	0.274	0.556
0.6	0.692	0.449	0.222	0.077	0.016	0.001	0.000	0.001	0.022	0.127	0.400
0.7	0.803	0.621	0.405	0.208	0.077	0.017	0.001	0.000	0.002	0.038	0.243
0.8	0.889	0.764	0.598	0.405	0.222	0.090	0.022	0.002	0.000	0.004	0.111
0.9	0.953	0.874	0.764	0.621	0.449	0.274	0.127	0.038	0.004	0.000	0.027
1	1.000	0.953	0.889	0.803	0.692	0.556	0.400	0.243	0.111	0.027	nd

$n = 100$	f2										
f1	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.0	nd	0.217	0.556	0.763	0.870	0.926	0.957	0.976	0.988	0.995	1.000
0.1	0.217	0.000	0.037	0.282	0.594	0.790	0.891	0.942	0.970	0.986	0.995
0.2	0.556	0.037	0.000	0.017	0.185	0.497	0.741	0.872	0.937	0.970	0.988
0.3	0.763	0.282	0.017	0.000	0.012	0.148	0.455	0.724	0.872	0.942	0.976
0.4	0.870	0.594	0.185	0.012	0.000	0.010	0.138	0.455	0.741	0.891	0.957
0.5	0.926	0.790	0.497	0.148	0.010	0.000	0.010	0.148	0.497	0.790	0.926
0.6	0.957	0.891	0.741	0.455	0.138	0.010	0.000	0.012	0.185	0.594	0.870
0.7	0.976	0.942	0.872	0.724	0.455	0.148	0.012	0.000	0.017	0.282	0.763
0.8	0.988	0.970	0.937	0.872	0.741	0.497	0.185	0.017	0.000	0.037	0.556
0.9	0.995	0.986	0.970	0.942	0.891	0.790	0.594	0.282	0.037	0.000	0.217
1	1.000	0.995	0.988	0.976	0.957	0.926	0.870	0.763	0.556	0.217	nd

Linkage disequilibrium within heterotic groups

Linkage disequilibrium within a group is related to the history of the group. In maize, major heterotic groups such as Reid Yellow Dent, Lancaster, and European Flint have been created from traditional population varieties. First-generation inbreds extracted from these populations were then intercrossed to generate second- and further generation inbreds (Hallauer 1990). Thus, linkage disequilibrium within a group could have appeared in: (1) the first-generation inbreds or (2) during the derivation of subsequent cycle inbreds. The disequilibrium in first-generation inbreds is founded upon disequilibrium in the source populations.

Linkage equilibrium of the source population is expected if mating has been nearly panmictic for many generations (cf. Charcosset et al. 1991). Experimental evidence indicates that natural allogamous populations are close to linkage equilibrium (Hastings 1989). Linkage disequilibrium in the source population is expected if the popualtion has been created recently through hybridization or has passed through a bottle-neck generated by severe selection pressure, or more likely, by random drift due to limited population size. If linkage is tight, bottle-neck effects can be maintained for a rather long period of time (Avery and Hill 1978). Significant linkage disequilibrium between isozyme loci in traditional maize populations was reported by Garnier-Géré (1992).

If the source population is in linkage equilibrium, disequilibrium between markers and QTLs can be randomly generated by a bottle-neck if the number of first-generation inbreds is small. If the source is in disequilibrium, disequilibrium between markers and QTLs in the set of first-generation lines is no longer generated at random but depends on the disequilibrium within the initial population.

Linkage disequilibrium can also arise during subsequent breeding after the first-generation inbreds have been derived. The history of the BSSS synthetic is illustrative. Smith (1983) and Helms et al. (1986) emphasized the importance of genetic drift over selection cycles. Helms et al. (1989) concluded that changes in allelic frequencies were due mostly to random drift. If so, linkage disequilibrium should have been generated concomitantly. In the case of European maize germ plasm, the European Flint heterotic group has been generated mainly from three first-generation inbreds (lines F2, F7, and EP1) and has undergone a maximum of four subsequent generations of line development. Linkage disequilibrium between isozyme alleles *Mdh5–15* and *Pgm2-a* is evident in the derived germ plasm (Bar-Hen et al. submitted). These two examples suggest that linkage disequilibria should generally exist within groups. If so, marker-based prediction of heterosis should generally be effective within heterotic groups.

Linkage disequilibria within different heterotic groups

If random effects play an important role in the origin of linkage disequilibrium, linkage disequilibrium between markers and QTLs should differ randomly among groups (Lewontin 1974). Striking differences between maize populations for disequilibrium between isozyme alleles have been reported by Garnier-Géré (1992). Similarly, Bar-Hen et al. (submitted) reported differences for the sign and the magnitude of the disequilibria between isozyme alleles in different groups of maize inbred lines. Thus, in accordance with formula 17, prediction of performance of hybrids between lines belonging to different groups on the basis of heterozygosity at neutral marker loci will not be effective. This conclusion is consistent with experimental results by Melchinger et al. (1992) and Boppenmeier et al. (1992).

The effect of groups allelic divergence

Allelic divergence among groups at the marker loci and the QTLs produces linkage disequilibrium between marker loci and QTLs involved in *SCA* (Eq. 19). This generates a correlation between *SCA* and heterozygosity at marker loci (Eq. 21) when considering simultaneously between-groups' and within-groups' hybrids. Effective exploitation of this phenomenon depends on: (1) prior knowledge of *SCA* among the groups and (2) the ability to determine group membership of parental inbreds.

Concerning (1), specific heterotic groups have been classified on the basis of between-groups' heterosis. Thus, these groups should differ for their allelic frequencies at the QTLs that exhibit dominance effects. The divergence should be accentuated between groups that have undergone reciprocal recurrent selection. Experimental evidence on the magnitude of *SCA* variation that can be accounted for by the partition of the inbreds into heterotic groups would aid in understanding the basis of the relationship between heterosis and heterozygosity at marker loci.

Concerning (2), differentiation of groups at marker loci has been reported by several authors (Godshalk et al. 1990; Messmer et al. 1992; Livini et al. 1992). All concurred positively regarding the possibility of using marker analysis to assign inbreds to heterotic groups. Group differentiation can be generated by random drift, as was discussed previously, or by reciprocal recurrent selection because of the linkage drag between markers and QTLs (i.e., hitchhiking effect).

The efficiency of the markers to assign inbreds to heterotic groups can be estimated through parameter Γ_m (see Eq. 24). Table 1 illustrates that, for given frequencies, Γ_m increases with the number of markers that are considered. However, differentiation of the groups depends on the specific markers that are considered. The parameter Γ_m could be used as a criterion to determine the optimum combination of markers that should be considered for distance computation.

Conclusions

Linkage disequilibrium between markers and QTLs should be generally expected in most populations. Thus, this necessary condition for prediction efficiency should be fulfilled at the within-group level and at the general level (provided a differentiation of the groups). However, the fact that linkage disequilibria between markers and QTLs generally differ randomly from one heterotic group to the other suggests that distances based on neutral marker loci will not be predictive of the performance of between-groups' hybrids, which is consistent with results from the studies of Melchinger et al. (1992) and Boppenmeier et al. (1992).

Several effective prediction methods may be possible. Heterozygosity at marker loci will be predictive only if linkage disequilibria between the markers and the QTLs are similar in the groups of interest. The mapping of QTLs involved in heterotic response (Stuber et al. 1992) should identify predictive markers. An alternative would be to use non-neutral genetic markers (Leonardi et al. 1991). Of course, probes specific to the genes involved in heterotic response are the ultimate solution.

Another possibility, which may be more readily applicable, could be to reconsider the statistical model that is used for prediction. At the within-group level, markers appear to be a powerful tool to estimate the genetic similarity between inbreds. Thus, if two lines that belong to the same heterotic group are close at the marker level, they should display similar *SCA* values (with testers of complementary groups). Thus, a possible scheme would be: (1) build a factorial design to determine the *SCA* values for a set of hybrids between lines that represent one group (set *Ref* 1) and lines that represent another group (set *Ref* 2); (2) use markers to estimate the similarity between the lines that belong to group 1 and the lines of set *Ref* 1, the similarity between the lines that belong to group 2 and the lines of set *Ref* 2; (3) develop an index function that includes *Ref* 1 × *Ref* 2 *SCA* estimates and similarity estimates to predict the *SCA* of Group 1 × Group 2 hybrids.

Appendix: derivation of *var*(SCA_{ij})

$var(SCA_{ij})$ can be written as a sum of expectations (over the hybrids):

$$var(SCA_{ij}) = E(SCA_{ij})^2$$

$$= \frac{1}{4}\sum_{k=1}^{nl}\sum_{l=1}^{nl} d_k d_l E((\theta_k^i - w_k)(w_k - \theta_k^j)(\theta_l^i - w_l)(w_l - \theta_l^j)) \quad (25)$$

Since we assumed that the diallel was complete, for any locus *l*, the two variables θ_l^i and θ_l^j, are independent. Thus we have:

$$var(SCA_{ij}) = \frac{1}{4}\sum_{k=1}^{nl}\sum_{l=1}^{nl} d_k d_l E^2((\theta_k^i - w_k)(\theta_l^i - w_l)) \quad (26)$$

$$var(SCA_{ij}) = \frac{1}{4}\sum_{k=1}^{nl} d_k^2 E^2((\theta_k^i - w_k)^2)$$

$$+\frac{1}{4}\sum_{k=1}^{nl}\sum_{l=1,l\neq k}^{nl} d_k d_l E^2((\theta_k^i - w_k)(\theta_l^i - w_l)) \tag{27}$$

$$var(SCA_{ij}) = \sum_{k=1}^{nl} d_k^2 H_k^2 + 4\sum_{k=1}^{nl}\sum_{l=1,l\neq k}^{nl} d_k d_l D_{lk}^2 \tag{28}$$

Acknowledgements We are grateful to A. Gallais for his helpful discussion and encouragement. We thank J. Burstin, D. de Vienne, F. Hospital, and an anonymous referee for helpful suggestions on the manuscript.

References

Avery PJ, Hill WG (1978) Distribution of linkage disequilibrium with selection and finite population size. Genet Res 33:29–45

Bar-Hen A, Charcosset A, Bourgoin M, Guiard J (submitted) Relationship between genetic markers and morphological traits in a maize inbred lines collection. Implications for breeding programs and owner's right protection

Beckmann JS, Soller M (1983) Restriction fragments length polymorphisms in genetic improvement: methodologies, mapping and costs. Theor Appl Genet 67:35–43

Bernardo R (1992) Relationship between single-cross performance and molecular marker heterozygosity. Theor Appl Genet 83:628–634

Boppenmeier J, Melchinger AE, Brunklaus-Jung E, Geiger HH, Herrmann RG (1992) Genetic diversity for RFLPs in European maize inbreds: I. relation to performance of Flint × Dent crosses for forage traits. Crop Sci 32:895–902

Burr B, Evola SV, Burr FA, Beckmann JS (1983) The application of restriction fragment length polymorphism to plant breeding. In: Setlow JK, Hollaender A (eds) Genetic engineering principle and methods, vol 5. Plenum Press, London, pp 45–59

Charcosset A (1992) Prediction of heterosis. In: Proc XIII Congr EUCARPIA. Springer, Berlin Heidelberg New York, 355–369

Charcosset A, Lefort-Buson M, Gallais A (1991) Relationship between heterosis and heterozygosity at marker loci: a theoretical computation. Theor Appl Genet 81:571–575

Crow JF, Kimura M (1970) An introduction to population genetics theory. Harper and Row, New York

Dudley JW, Sagai Maroof MA, Rufener GK (1991) Molecular markers and grouping of parents in maize breeding programs. Crop Sci 31:718–723

Falconer DS (1981) An introduction to quantitative genetics, 2nd edn. Longman, London

Frei OM, Stuber CW, Goodman MM (1986) Uses of allozymes as genetic markers for predicting performance in maize single-cross hybrids. Crop Sci 26:37–42

Gardner CO, Eberhart SA (1966) Analysis and interpretation of the variety cross diallel and related populations. Biometrics 22: 439–452

Garnier-Géré P (1992) Contribution à l'étude de la variabilité génétique inter- et intra- population chez le maïs (*Zea mays* L.): valorisation d'informations agromorphologiques et enzymatiques. PHD thesis Institut National Agronomique Paris Grignon

Godshalk EB, Lee M, Lamkey KR (1990) Relationship of restriction length polymorphisms to single-cross hybrid performance of Maize. Theor Appl Genet 80:273–280

Griffing B (1956) Concept of general and specific combining ability in relation to diallel crossing system. Aust J Biol Sci 9:463–493

Hallauer AR (1990) Methods used in developing maize inbreds. Maydica 35:1–17

Hallauer AR, Miranda FO (1988) Quantitative genetics in maize breeding. Iowa State University Press, Ames, Iowa

Hastings A (1989) The interaction between selection and linkage in plant populations. In: Brown AHD, Clegg MT, Kahler AL, Weir BS (eds) Plant population genetics, breeding, and genetic resources. Sinaver, Boston, Mass., pp 163–180

Hayman BI (1954) The theory and analysis of diallel crosses. Genetics 39:789–809

Helms TC, Hallauer AR, Smith OS (1989) Genetic drift and selection evaluated from recurrent selection programs in maize. Crop Sci 29:602–607

Lee M, Godshalk EB, Lamkey KR, Woodman WL (1989) Association of restriction length polymorphisms among maize inbreds with agronomic performance of their crosses. Crop Sci 29:1067–1071

Leonardi A, Damerval C, Hébert Y, Gallais A, de Vienne D (1991) Association of protein amount polymorphism (PAP) among maize lines with performances of their hybrids. Theor Appl Genet 82:552–560

Lewontin RC (1974) The genetic basis of evolutionary change. Colombia University Press, New York London

Livini C, Ajmone-Marsan P, Melchinger AE, Messmer MM, Motto M (1992) Genetic diversity of maize inbred lines within and among heterotic groups revealed by RFLPs. Theor Appl Genet 84:17–25

Melchinger AE, Lee M, Lamkey KR, Hallauer AR, Woodman WL (1990a) Genetic diversity for restriction fragment length polymorphisms and heterosis for two diallel sets of maize inbreds. Theor Appl Genet 80:488–496

Melchinger AE, Lee M, Lamkey KR, Woodman WL (1990b) Genetic diversity for restriction fragment length polymorphisms: relation to estimated genetic effects in maize inbreds. Crop Sci 30: 1033–1040

Melchinger AE, Boppenmeier J, Dhillon BS, Pollmer WG, Herrmann RG (1992) Genetic diversity for RFLPs in European maize inbreds: II. relation to performance of hybrids within versus between heterotic groups for forage traits. Theor Appl Genet 84:672–681

Messmer MM, Melchinger AE, Boppenmeier J, Herrmann RG, Brunklaus-Jung E (1992) RFLP analyses of early-maturing European maize germ plasm. I. Genetic diversity among flint and dent inbreds. Theor Appl Genet 83:1003–1012

Moll RH, Lonnquist JH, Fortuna JV, Johnson EC (1965) The relationship of heterosis and genetic divergence in maize. Genetics 52:139–144

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 94:3321–3323

Otha T (1982) Linkage disequilibrium due to random drift in finite subdivided populations. Proc Natl Acad Sci USA 79:1940–1944

Rogers JS (1972) Measures of genetic similarity and genetic distance. In: Wheeler MR (ed) Studies in genetics VII. University of Texas Publ 7213, Houston, Tex., pp 145–153

Roughgarden (1979) Theory of population genetics and evolutionary ecology: an introduction. MacMillan Publ, New York

Shull GH (1908) The composition of a field of maize. Am Breed Assoc Rep 4:296–301

Smith OS (1983) Evaluation of recurrent selection in BSS, BSCB1, and BS13 maize population. Crop Sci 23:35–40

Smith OS, Smith JSC, Bowen SL, Tenborg RA, Wall SJ (1990) Similarities among a group of elite maize inbreds as measured by pedigree, F_1 grain yield, grain yield heterosis, and RFLPs. Theor Appl Genet 80:833–840

Sprague GF, Tatum LA (1942) General vs. specific combining ability in single crosses of corn. J Am Soc Agron 34:923–932

Stuber CW (1989) Marker-based selection for quantitative traits. In: G. Robbelen (ed) Proc XII Congr EUCARPIA. Parey, Berlin Hamburg, pp 31–49

Stuber CW, Lincoln SE, Wolff DW, Helentjaris T, Lander ES (1992) Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. Genetics 132:823–839