Heredity **79** (1997) 477–483 *Received 26 October 1996*

Relationship between phenotypic and marker distances: theoretical and experimental investigations

JUDITH BURSTIN & ALAIN CHARCOSSET*

INRA-UPS-INAPG, Station de Génétique Végétale, Ferme du Moulon, 91190 Gif/Yvette, France

Numerous studies have aimed at assessing the relationships between (i) distances computed from phenotypic data, (ii) distances computed from marker data and (iii) heterosis, for pairs of individuals or populations. The conflicting results obtained illustrate that these relationships are far from simple. In this paper, we investigate the effect on these relationships of (i) the polygenic inheritance of phenotypic traits and (ii) the structure of linkage disequilibrium between genetic markers and the loci involved in the variation of quantitative traits (QTLs). Both theoretical and experimental results showed that the relationship between marker distances and phenotypic distances computed from quantitative traits displays a triangular shape: low marker distances are systematically associated with low phenotypic distances, whereas high marker distances correspond to either low or high phenotypic distances. Because of this property, the linear coefficient of correlation between both distances decreases as the number of QTLs involved in the variation of the traits considered for phenotypic distance computation increases. Similar properties are expected for the relationship between heterosis and phenotypic distances.

Keywords: genetic distances, heterozygosity, markers, polygenic inheritance.

Introduction

Genetic differentiation between individuals or populations can be evaluated at different levels: quantitatively inherited phenotypic traits, monogenic traits submitted to selection pressure (e.g. disease resistance traits), neutral molecular markers, etc. Information about the relationships that exist between these different levels is significant for several reasons. From an evolutionary standpoint, the investigation of the relationship between genetic diversity and morphological differentiation has been expected to give clues to the forces that are possibly responsible for this differentiation. From a genetic resources conservation point of view, it may be useful to know whether or not two individuals or populations that are phenotypically similar display similar gene combinations. From an applied breeding point of view, phenotypic or genetic distances have been expected to provide predictors for heterosis.

*Correspondence. E-mail: charcos@moulon.inra.fr

Based on these reasons, the relationship between diversity at marker loci and morphological differentiation has been investigated in several studies. Some significant correlations between the two distances were reported (Atchley *et al.*, 1988), but in most cases no significant correlation was found (Wayne & O'Brien, 1986; Moser & Lee, 1994; Schmitt *et al.*, 1995). Similarly, very few experimental studies have demonstrated a relationship between heterosis (which can be considered as a particular distance; see Falconer, 1981) and distance parameters based on quantitative variations of phenotypic traits (Lefort-Buson, 1985). This lack of a clear relationship has been interpreted as resulting from irrelevant choices of the phenotypic traits taken into account for the calculation of the quantitative distance (Siiddiqui *et al.*, 1977; Partap *et al.*, 1980) or of the genotypes crossed (Peter & Rai, 1978), or from bad appreciations of the genotype as related to genotype × environment interaction (Singh & Ramanujam, 1981; Ghaderi *et al.*, 1984). More recently, the need for a linkage disequilibrium between the genes involved in the calculation of the

different distances and in heterosis has been emphasized (Charcosset *et al.*, 1991; Charcosset & Essioux, 1994; Burstin *et al.*, 1995).

The aim of this paper is to consider the effect of the polygenic inheritance of the traits used to compute phenotypic distances on the relationship between these distances and heterosis or marker distances. We show that the relationship between distances computed from quantitative phenotypic trait(s) and distances computed at individual loci or heterosis is not a linear one and discuss the parameters that affect the magnitude of the correlation. This is exemplified by experimental results on the relationship between distances computed from protein quantitative variations revealed by two-dimensional electrophoresis and a genetic distance computed from marker polymorphism among 21 maize inbred lines. The choice of protein quantities as the phenotypic characters used to estimate quantitative distances was based on major advantages over morphological traits: (i) they provide a high number of quantitative traits (190 in the present study), (ii) the complexity of genetic mechanisms involved in their variation ranges from oligogenic to polygenic and (iii) they are not affected by environmental fluctuations because they are assessed on 8-day-old seedlings grown under controlled conditions.

Theory

Basis of the model

Quantitative traits. We consider a biallelic model describing the phenotypic value of homozygous inbred lines, following the notations used in Charcosset *et al.* (1991). The genotype of inbred line i at locus l (with alleles l_1 and l_2) is represented by the variable θ_l^i, which takes the value $+1$, -1 for genotypes l_1l_1 and l_2l_2, respectively. The single-locus model for the phenotypic value of inbred line i is written as:

$$Y_i = c_l + a_l\theta_l^i,$$

where c_l is the average value of homozygotes l_1l_1 and l_2l_2, and a_l is half the difference between homozygous l_1l_1 and l_2l_2 phenotypes. If the trait is controlled by n_l loci acting independently (no epistasis), the phenotype of inbred line i (Y_i) is (with $C = \Sigma_{l=1}^{n_l} c_l$):

$$Y_i = C + \sum_{l=1}^{n_l} a_l\theta_l^i. \tag{1}$$

The phenotypic distance between i and j was defined as:

$$R_{ij}^2 = (Y_i - Y_j)^2 \tag{2}$$

or:

$$R_{ij} = |Y_i - Y_j|. \tag{3}$$

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

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

As emphasized by Charcosset & Essioux (1994), this model is also adapted to the case where more than two alleles are detected at marker loci [which is a general case for restriction fragment length polymorphisms (RFLPs) in many species].

Heterosis The heterosis expressed in the cross between inbred lines i and j is related to the heterozygosity of the hybrid $i \times j$ at the QTLs that display dominance effects (see Charcosset *et al.*, 1991, for a formal expression). Results presented further on for the relationship between quantitative distances and marker distances can therefore be readily extrapolated to the relationship between quantitative distance and heterosis if all QTLs involved in heterosis have dominance effects of the same magnitude.

Reference population We assume that the homozygous inbred lines belong to a reference population (i.e. a set of lines of infinite size). We define w_l as the mean of θ_l^i in this population. The frequency of the allele l_1 in the population is: $f_{l_1} = (1 + w_l)/2$. Genetic diversity at locus l (H_l) is proportional to the variance of θ_l^i: $H_l = \mathrm{Var}(\theta_l^i)/2 = (1 - w_l^2)/2$. In this paper, the linkage disequilibrium between two loci l and k is supposed to be either null or maximal (so that in this last situation, $\theta_l^i = \theta_k^i$ for all inbred lines i, or $\theta_l^i = -\theta_k^i$, for all inbred lines i).

In particular, we consider that (i) marker loci are independent, (ii) QTLs are independent, and (iii) among the n_p marker loci and the n_l QTLs, n_{lp} loci are common to both sets, i.e. being markers and QTLs.

Analytical results

Under previous hypotheses concerning the magnitude of linkage disequilibrium:

$$\mathrm{Var}(MD_{ij}) = \frac{1}{4n_p^2}\sum_{p=1}^{n_p}(1-w_p^4) \quad (5)$$

(Charcosset *et al.*, 1991). For simplicity, the calculation of the variance of phenotypic distance was only performed for R_{ij}^2. It can been demonstrated (see Appendix) that:

$$\mathrm{Var}(R_{ij}^2) = 4\sum_{l=1}^{n_l} a_l^4(1-w_l^4) + 8\sum_{l=1}^{n_l}\sum_{l'\neq l}^{n_l} a_l^2 a_{l'}^2(1-w_l^2)(1-w_{l'}^2) \quad (6)$$

and

$$\mathrm{Cov}(R_{ij}^2, MD_{ij}) = \frac{1}{n_p}\sum_{k=1}^{n_{lp}} a_k^2(1-w_k^4). \quad (7)$$

We have considered in particular the case of equal allelic frequencies at each locus (for all loci l, $f_{l_1} = f_{l_2} = 0.5$), further considering that all QTLs have the same contribution to the variation of the trait of interest (for all loci l, $a_l = a$). In this case, the correlation coefficient between phenotypic and marker distances is:

$$\rho(R_{ij}^2, MD_{ij}) = \frac{n_{lp}}{\sqrt{n_p}\sqrt{2n_l^2 - n_l}} \quad (8)$$

$$= \frac{n_{lp}}{\sqrt{n_p}\sqrt{n_l}} \times \frac{1}{\sqrt{2n_l - 1}} \quad (9)$$

This last expression illustrates that the magnitude of $\rho(R_{ij}^2, MD_{ij})$ depends on two factors: (i) the association between the marker loci and the QTLs ($n_{lp}/\sqrt{n_p}\sqrt{n_l}$), and (ii) the number of loci involved in the variation of the quantitative trait ($1/\sqrt{2n_l - 1}$).

Numerical results

According to the assumptions used for the analytical developments, we considered n_l loci involved with equal contributions in the variation of the trait of interest, each locus being biallelic with equal frequencies (0.5) for the two alleles. We further considered that $n_l = n_p = n_{lp}$. For $n_{lp} = 2$ to $n_{lp} = 8$, we computed distance parameters (MD_{ij}, R_{ij}, R_{ij}^2, defined as previously) between the $2^{n_{lp}} \times 2^{n_{lp}}$ possible pairs of genotypes at the n_{lp} loci.

Table 1 Correlation between marker distance (MD) and quantitative distances R and R^2: $\rho(MD, R)$ and $\rho(MD, R^2)$, for n_{lp} loci (see text for hypotheses). N indicates the number of pairs of genotypes involved in the correlation computation

n_{lp}	2	3	4	5	6	7	8
N	16	64	256	1024	4096	16384	65536
$\rho(MD, R)$	0.53	0.41	0.35	0.31	0.28	0.26	0.24
$\rho(MD, R^2)$	0.58	0.45	0.38	0.33	0.30	0.28	0.26

We checked that under these conditions the numerical values for the correlation between MD_{ij} and R_{ij}^2 were consistent with those obtained from the analytical approach (Table 1). The correlation between MD_{ij} and R_{ij} was slightly lower than between MD_{ij} and R_{ij}^2. The correlation between the marker distance MD_{ij} and the quantitative distance parameters R_{ij} and R_{ij}^2 decreased as the number of loci considered increased (Table 1). For $n_{lp} = 2$ to 8, triangular relationships were observed (Fig. 1): low MD_{ij} values were systematically associated with low

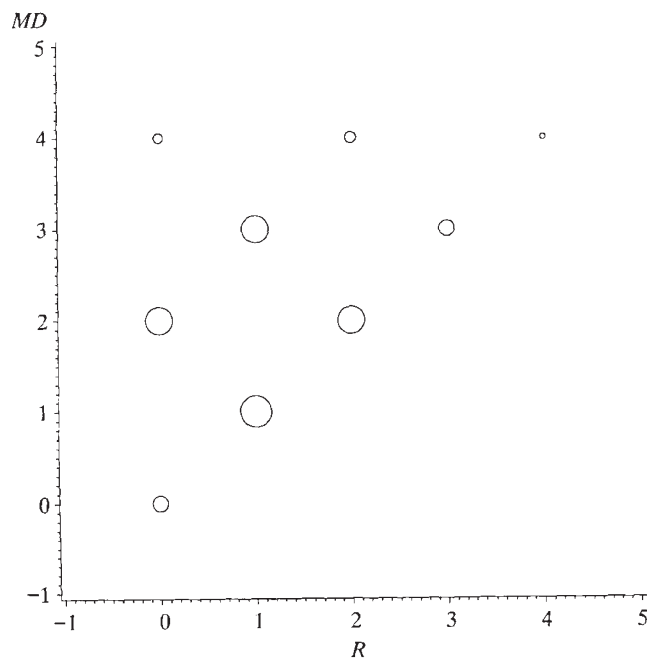

Fig. 1 Marker distance (MD, ordinate) vs. quantitative distance (R, abscissa) for a four-loci model (see text for model hypotheses). The size of each circle is proportional to its frequency.

quantitative distances, whereas high MD_{ij} values corresponded to either low or high quantitative distance. With equal and independent effects of the increasing and the decreasing alleles at the different loci, two different genotypes, for example $(++--)$ and $(--++)$, can have the same phenotype although they are different at each locus, whereas identical genotypes necessarily have the same phenotype.

Materials and methods

Twenty-one maize inbred lines have been characterized as described in Burstin *et al.* (1994) for 142 markers resulting from the analysis of enzyme, RFLP and anonymous protein polymorphisms, and for the relative quantities of 190 proteins revealed by two-dimensional electrophoresis. The protein quantities were determined by the KEPLER 2-D Gel analysis Software, as described in Burstin *et al.* (1993). For each pair of lines, we computed the multilocus Rogers's distance for the 142 marker loci (*MD*), and a distance defined for each of the 190 proteins as follows: $Rk_{ij} = (Q_{ik} - Q_{jk})^2$ with Q_{ik} and Q_{jk} the standardized quantities of protein k in lines i and j, respectively.

Experimental results

As expected, the correlations between *MD* and the 190 phenotypic distances corresponding to the 190 protein quantities were generally small. They ranged from -0.24 to 0.48, with 97 per cent of the correlation coefficients being between -0.25 and 0.25. The highest correlation coefficient (0.48) corresponded to the protein S207, which variation can be considered as oligogenic because QTL analysis demonstrated that it is mostly determined by three loci (Damerval *et al.*, 1994). This protein has been identified by microsequence comparisons as a gluthatione-S-transferase (Touzet *et al.*, 1995). Moreover, a trend towards a triangular relationship clearly appeared on the plots of *MD* vs. the phenotypic distances computed on the standardized phenotypic of proteins. For example, Fig. 2 shows the relationship observed between *MD* and the phenotypic distance computed from the variation of protein S65. This protein has been identified by amino acid composition comparisons as a chaperonin hsp60 (Touzet *et al.*, 1996). Consistently with numerical results, low *MD* values were associated with low phenotypic distances, whereas high *MD* values corresponded to either low or high quantitative distance.

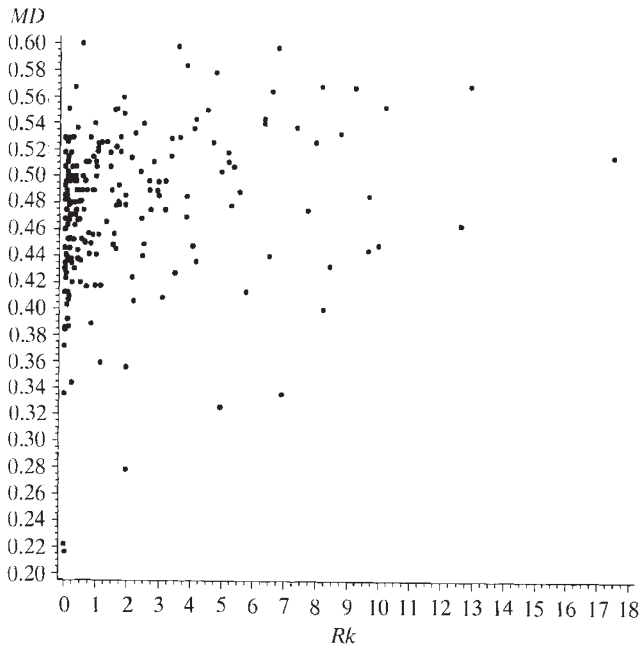

Fig. 2 Relationship between a marker distance computed from 142 marker loci (*MD*, ordinate) and a quantitative distance computed for protein S65 (*Rk*, abscissa).

Discussion

Experimental results reported in this study underlined that, for a number of proteins, low marker distances were associated with low phenotypic distances, whereas high marker distances were associated with a large range of phenotypic distances. Thus, a clear tendency towards a triangular relationship was observed between the two distances (Fig. 2). This result is consistent with other studies. Burstin *et al.* (1995) found a similar relationship between a marker distance and Hanson and Casas distances computed for yield and early vigour on 210 pairs of maize inbred lines. A triangular relationship was also found between a marker distance computed from 222 marker loci and a Mahalanobis distance computed on 10 phenotypic traits, for 10 440 pairs of lines (Bar-Hen & Charcosset, 1995; Dillmann *et al.*, 1997). Similar results were also reported by Chantereau (1993) for sorghum inbreds. Triangular relationships were also observed between heterosis and phenotypic distances (Charcosset *et al.*, 1990, and our unpublished data). However, Leonardi *et al.* (1991) reported a linear relationship between the protein quantitative distance and heterosis for five agromorphological characters in maize. This result could be related to (i) the smaller number of inbred lines that were considered in that study (five vs. 21 in the present study), which may have led to sampl-

ing problems, and (ii) different scoring methods for the protein quantities (discussed in Burstin *et al.*, 1995).

The triangular shape of the relationship between the marker distance or heterosis and the distance computed for phenotypic traits can be explained by the polygenic inheritance of these traits, because a given quantitative value can be obtained with different gene combinations. Polygenic inheritance has been demonstrated for most of the traits generally considered for phenotypic distance estimation. For example, QTL mapping in maize has revealed that numerous chromosome regions are involved in plant height variation (e.g. Beavis *et al.*, 1991). Damerval *et al.* (1994) have demonstrated that individual protein quantity variations were often polygenically inherited in an F_2 progeny of maize. Thus, the number of QTLs involved in the variation of a phenotypic trait in a sample of diverse genotypes is generally expected to be large. Our theoretical results demonstrate that the correlation between phenotypic distances and marker distances or heterosis necessarily decreases with the number of loci involved in the variation of the trait(s) of interest. This result is consistent with the small correlations observed between *MD* and the distances computed from protein quantities, and the fact that the highest correlation is observed for a protein with quantity controlled by a restricted number of loci. The polygenic inheritance of the quantitative traits taken into account in the phenotypic distance computation has a similar influence on the relationship between the distance and heterosis.

In addition to the polygenic inheritance of quantitative traits, one has to consider that markers generally have no direct effect on the quantitative traits of interest (i.e. they are neutral). Thus, the relationship between quantitative distances and marker distances is affected by the linkage disequilibrium between marker loci and the QTLs involved in the traits considered for quantitative distance estimation. Equation (9) illustrates that a poor association between both types of loci leads to a low correlation between distances. If there is no linkage disequilibrium, the two distances vary independently: high and low marker distances can correspond to similar morphological distances. If there is linkage disequilibrium, which is, for example, the case when inbred lines are related by pedigree (Charcosset and Essioux, 1994), a strong relationship is expected. Because a high marker similarity is necessarily associated with kinship (e.g. Smith *et al.*, 1990), two lines that are similar at marker loci will share common alleles at the QTLs and thus be phenotypically close. This effect of kinship on the triangular shape of the relationship was illustrated experimentally in the present study: pairs of related lines corresponded to small distances, of either *MD* or *R*, whereas pairs of unrelated lines corresponded to large values of *MD* and a large range of values for *R*. Thus, polygenic inheritance and linkage disequilibrium properties associated with kinship lead to a triangular relationship between marker distance and phenotypic distance. The relationship between phenotypic distances and heterosis depends in a similar way on the linkage disequilibrium between QTLs involved in heterosis and QTLs involved in the traits considered for phenotypic distance estimation.

The results reported in this study illustrate that prediction of heterosis based on quantitative distances between parents has to be considered with caution. As also discussed in Charcosset *et al.* (1990), a hybrid combination between two parents should not be discarded a priori on the basis of their morphological similarity, because similar phenotypes can be observed for different genetic combinations. The relationship between marker and phenotypic distances also has several practical applications. First, the phenotype of a given line can be predicted if it displays a high similarity at the marker level with a line that was characterized phenotypically. Secondly, among a set of lines with similar phenotypes, marker analysis allows the identification of the lines that are likely to share similar alleles at the QTLs. This can be extremely interesting for the protection of owner's rights as well as for genetic resources conservation. However, because of the linkage disequilibrium properties, two inbreds may have similar phenotypes, share common alleles at the QTLs and display a relatively high marker distance at the same time. These applications would be broadened by the identification of the genes involved in the variation of the traits of interest, which would allow an allelic comparison at the QTL level.

Acknowledgements

We are grateful to C. Damerval, P. Dubreuil, A. Gallais, M. Lefort and D. de Vienne for helpful discussions, to P. Dubreuil and P. Dufour for their contribution to RFLP results and to M. Grenèche for isozymic results.

References

ATCHLEY, W. R., NEWMAN, S. AND COWLEY, D. E. 1988. Genetic divergence in mandible form in relation to

molecular divergence in inbred mouse strains. *Genetics*, **120**, 239–253.

BAR-HEN, A. AND CHARCOSSET, A. 1995. Relationship between molecular and morphological distances in a maize inbred lines collection – application for breeders' rights protection. In: van Ooijen, J. W. and Jansen, J. (eds) *Biometrics in Plant Breeding: Applications of Molecular Markers*, pp. 57–66. Proceedings of the Ninth Meeting of the EUCARPIA Section Biometrics in Plant Breeding, Wageningen.

BEAVIS, W. D., GRANT, D., ALBERTSEN, M. AND FINCHER, M. 1991. Quantitative trait loci for plant height in four maize populations and their associations with qualitative genetic loci. *Theor. Appl. Genet.*, **83**, 141–145.

BURSTIN, J., ZIVY, M., DE VIENNE, D. AND DAMERVAL, C. 1993. Analysis of scaling methods to minimize experimental variations in two-dimensional electrophoresis quantitative data. Applications to the comparison of maize inbred lines. *Electrophoresis*, **14**, 1067–1073.

BURSTIN, J., DE VIENNE, D., DUBREUIL, P. AND DAMERVAL, C. 1994. Molecular markers and protein quantities as genetic descriptors in maize. I. Genetic diversity among 21 maize inbred lines. *Theor. Appl. Genet.*, **89**, 943–950.

BURSTIN, J., CHARCOSSET, A., BARRIERE, Y., HEBERT, Y., DE VIENNE, D. AND DAMERVAL, C. 1995. Molecular markers and protein quantities as genetic descriptors in maize. II. Prediction of performance of hybrids for forage traits. *Pl. Breed.*, **114**, 427–433.

CHANTEREAU, J. 1993. *Etude de l'hétérosis chez le sorgho (*Sorghum bicolor *L. Moench) par l'exploitation d'écotypes et l'analyse de leurs divergences*. Ph.D. Thesis, Université Paris Sud.

CHARCOSSET, A. AND ESSIOUX, L. 1994. The effect of population structure on the relationship between heterosis and heterozygosity at marker loci. *Theor. Appl. Genet.*, **89**, 336–343.

CHARCOSSET, A., LEFORT-BUSON, M. AND GALLAIS, A. 1990. Use of top-cross designs for predicting performance of maize single cross hybrids. *Maydica*, **35**, 23–27.

CHARCOSSET, A., LEFORT-BUSON, M. AND GALLAIS, A. 1991. Relationship between heterosis and heterozygosity at marker loci: a theoretical computation. *Theor. Appl. Genet.*, **89**, 571–575.

DAMERVAL, C., MAURICE, A., JOSSE, J. M. AND DE VIENNE, D. 1994. Quantitative trait loci underlying gene product variation – a novel perspective for analysing regulation of genome expression. *Genetics*, **137**, 289–301.

DILLMANN, C., BAR-HEN, A., GUERIN, D., CHARCOSSET, A. AND MURIGNEUX, A. 1997. Comparison of RFLP and morphological distances between maize *Zea mays* L. inbred lines. Consequences for germplasm protection purposes. *Theor. Appl. Genet.* (in press).

FALCONER, D. S. 1981. *Introduction to Quantitative Genetics*, 2nd edn. Longman, London.

GHADERI, A., ADAMS, M. W. AND NASSIB, A. M. 1984. Relationship between genetic distance and heterosis for yield and morphological traits in dry edible bean and faba bean. *Crop Sci.*, **24**, 37–42.

LEFORT-BUSON, M. 1985. Distance génétique et hétérosis 4. Utilisation de critères biométriques. In: Lefort-Buson, M. and de Vienne, D. (eds) *Les Distances Génétiques*. INRA, Paris. pp. 143–157.

LEONARDI, A., DAMERVAL, C., HEBERT, Y., GALLAIS, A. AND DE VIENNE, D. 1991. Association of protein amount polymorphism (PAP) among maize lines with performance of their hybrids. *Theor. Appl. Genet.*, **82**, 552–560.

MOSER, H. AND LEE, M. 1994. RFLP variation and genealogical distance, multivariate distance, heterosis, and genetic variance in oats. *Theor. Appl. Genet.*, **87**, 947–956.

PARTAP, P. S., DHANKHAR, B. S., PANDITA, M. L. AND DUDI, B. S. 1980. Genetic divergence in parents and their hybrids in Okra (*Abelmoschus esculentus* (L.) Moench). *Genet. Agr.*, **34**, 323–330.

PETER, K. V. AND RAI, B. 1978. Heterosis as a function of genetic distance in tomato. *Indian J. Genet. Plant Breed.*, **38**, 173–178.

ROGERS, J. S. 1972. Measures of genetic similarity and genetic distance. In: Wheeler, M. R. (ed.) *Studies in Genetics VII*, pp. 145–173. University of Texas Publ. 7213.

SCHMITT, L. H., KITCHENER, D. J. AND HOW, R. A. 1995. A genetic perspective of mammalian variation and evolution in the Indonesian archipelago: biogeographic correlates in the fruit bat genus *Cynopterus*. *Evolution*, **49**, 399–412.

SIDDIQUI, J. A., PRASAD, R. C. AND MEHRAR, B. 1976. Hybrid performance in relation to genetic divergence in some varieties of *Gossypium hirsutum*. *Pflanzenzucht.*, **77**, 215–221.

SINGH, S. P. AND RAMANUJAM, S. 1981. Genetic divergence and hybrid performance in *Cicer arietinum* L. *Indian J. Genet.*, **41**, 268–276.

SMITH, O. S., SMITH, J. S. C., BOWEN, S. L., TENBORG, R. A. AND WALL, S. J. 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.

TOUZET, P., MORIN, C., DAMERVAL, C., LE GUILLOUX, M., ZIVY, M. AND DE VIENNE, D. 1995. Characterizing allelic proteins for genome mapping in maize. *Electrophoresis*, **16**, 1289–1294.

TOUZET, P., DE VIENNE, D., HUET, J. C., OUALI, C., BOUET, F. AND ZIVY, M. 1996. Amino acid analysis of proteins separated by two-dimensional electrophoresis in maize: Isoform detection and function identification. *Electrophoresis*, **17**, 1393–1401.

WAYNE, R. K. AND O'BRIEN, S. J. 1986. Empirical demonstration that structural genes and morphometric variation of mandible traits are uncoupled between mouse strains. *J. Mammal.*, **67**, 441–449.

Appendix

Variance of R^2

$$\mathrm{Var}(R_{ij}^2) = \mathrm{Var}\left(\left(\sum_l a_l(\theta_l^j-\theta_l^i)\right)^2\right)$$

$$= \mathrm{Cov}\left(\sum_l\sum_{l'} a_l a_{l'}(\theta_l^j-\theta_l^i)(\theta_{l'}^j-\theta_{l'}^i);\sum_k\sum_{k'} a_k a_{k'}(\theta_k^j-\theta_k^i)(\theta_{k'}^j-\theta_{k'}^i)\right)$$

$$= \sum_l\sum_{l'}\sum_k\sum_{k'} a_l a_{l'} a_k a_{k'}\alpha_{ll'kk'},$$

where $\alpha_{ll'kk'} = \mathrm{Cov}((\theta_l^j-\theta_l^i)(\theta_{l'}^j-\theta_{l'}^i);\ (\theta_k^j-\theta_k^i)(\theta_{k'}^j-\theta_{k'}^i))$.

Based on the elementary covariance property ($\mathrm{Cov}(AB;\ C) = \mathrm{E}(A)\mathrm{Cov}(B;C)$) if A is independent of B and C) and the specific property ($\mathrm{E}(\theta_l^j-\theta_l^i) = 0$)), $\alpha_{ll'kk'} = 0$ if one locus is independent from the three other loci. In addition, $\alpha_{ll'kk'} = 0$ if $l = l'$, $k = k'$ and l and k are independent. If one further assumes that the linkage disequilibrium between two loci can only be total or null, $\alpha_{ll'kk'} \neq 0$ in the two following situations.

- $l = k$, $l' = k'$ and $l \neq l'$ (or $l = k'$, $l' = k$ and $l \neq l'$). In this situation,

$$\alpha_{ll'kk'} = \mathrm{E}((\theta_l^j-\theta_l^i)^2(\theta_k^j-\theta_k^i)^2)-\mathrm{E}^2((\theta_l^j-\theta_l^i)(\theta_k^j-\theta_k^i))$$
$$= \mathrm{E}((2-2\theta_l^i\theta_l^j)(2-2\theta_k^i\theta_k^j))-0$$
$$= 4(1-w_l^2)(1-w_k^2).$$

- $l = k = l' = k'$. In this situation,

$$\alpha_{ll'kk'} = \mathrm{Cov}(-2\theta_l^i\theta_l^j;\ -2\theta_l^i\theta_l^j)$$
$$= 4(\mathrm{E}(\theta_l^i\theta_l^j)^2)-\mathrm{E}^2(\theta_l^i\theta_l^j)$$
$$= 4(1-w_l^4).$$

Thus:

$$\mathrm{Var}(R_{ij}^2) = 4\sum_l a_l^4(1-w_l^4)+8\sum_l\sum_{k\neq l} a_l^2 a_k^2(1-w_l^2)(1-w_k^2).$$

Covariance between R^2 *and* MD

$$\mathrm{Cov}(R_{ij}^2;\ MD_{ij}) = \mathrm{Cov}\left(\left(\sum_l a_l(\theta_l^j-\theta_l^i)\right)^2;\ \frac{1}{2n_p}\sum_k(1-\theta_k^i\theta_k^j)\right)$$

$$= \frac{1}{2n_p}\mathrm{Cov}\left(\sum_l\sum_{l'} a_l a_{l'}(\theta_l^j-\theta_l^i)(\theta_{l'}^j-\theta_{l'}^i);\ \sum_k(1-\theta_k^i\theta_k^j)\right)$$

$$= \frac{1}{2n_p}\sum_l\sum_{l'}\sum_k a_l a_{l'}\mathrm{Cov}((\theta_l^j-\theta_l^i)(\theta_{l'}^j-\theta_{l'}^i);\ (1-\theta_k^i\theta_k^j)).$$

Following previous assumptions concerning the possible magnitude of linkage disequilibrium, $\mathrm{Cov}((\theta_l^j-\theta_l^i)(\theta_{l'}^j-\theta_{l'}^i);\ (1-\theta_k^i\theta_k^j)) = 0$, unless $l = l' = k$. For this last situation, it can be noted that $(\theta_l^j-\theta_l^i)^2 = 2(1-\theta_l^i\theta_l^j)$ and so: $\mathrm{Cov}((\theta_l^j-\theta_l^i)^2;\ (1-\theta_l^i\theta_l^j)) = 2\mathrm{Var}(1-\theta_l^i\theta_l^j)$. So,

$$\mathrm{Cov}(R_{ij}^2;\ MD_{ij}) = \frac{1}{n_p}\sum_{l=1}^{n_{ip}} a_l^2(1-w_l^4).$$