Theor Appl Genet (1997) 95:92–102 © Springer-Verlag 1997

C. Dillmann · A. Bar-Hen · D. Guérin
A. Charcosset · A. Murigneux

Comparison of RFLP and morphological distances between maize *Zea mays* L. inbred lines. Consequences for germplasm protection purposes

Received: 20 August 1996 / Accepted: 20 December 1996

Abstract A total of 145 maize inbred lines, representative of material released in France, were differentiated using RFLP markers and a set of discriminant morphological traits in order to evaluate the use of molecular markers for large-scale germplasm diversity analysis and determination of distinctness. Several criteria are proposed with respect to choice of probes, which should give reliable results for routine studies and have a known single-locus genetic determinism to avoid redundancy. A method is proposed by which to incorporate the data from different restriction enzymes obtained with the same probe. The precision of the estimation of the genetic distance is given. The relationship between molecular and morphological distances appears to be triangular, molecular divergence behaving as a limiting factor for morphological divergence. This suggested a scheme for incorporating molecular markers in studies of distinctness.

Key words RFLP · Markers · Genetic distances · Morphological distances

Communicated by P. L. Pfahler

C. Dillmann[1](✉) · A. Bar-Hen · D. Guérin
GEVES, La Minière, F-78280 Guyancourt Cedex, France

A. Charcosset
INRA Station de Génétique Végétale, Ferme du Moulon, F-91190 Gif sur Yvette, France

A. Murigneux
Laboratorie BIOCEM, 24 Avenue des Landais, F-63170 Aubière, France

Present address:
[1] INRA Station de Génétique Végétale, Ferme du Moulon, F-91190 Gif sur Yvette, France

Introduction

The accurate description of new varieties is important to allow their inscription or protection through plant variety protection systems. UPOV (Union pour la Protection des Obtentions Végétales) guidelines stipulate that a newly released variety must be distinctly different from all previously released varieties. Comparisons between cultivars are currently performed on the basis of a large set of morphological traits recorded at several stages of plant growth and chosen for their discriminational power (see Smith and Smith 1989 for maize). Molecular markers have always been viewed as additional tools for varietal description (Soller and Beckmann 1983), and their discriminational power has been extensively studied in maize (Smith et al. 1990; Smith et al. 1991b; Melchinger et al. 1991; Messmer et al. 1991; Bernardo 1993; Dubreuil et al. 1996) as well as their relationship with yield and heterosis (Lee et al. 1989; Godshalk et al. 1990; Melchinger et al. 1990; Dudley et al. 1991; Burstin et al. 1997).

The potential use of molecular markers for distinguishing between maize inbred lines has been discussed by Smith et al. (1991a). It clearly depends on four parameters: (1) the quality of the molecular markers; (2) the choice of a distance index which suits the molecular data and the specific plant material; (3) the precision of the estimation of the genetic distance, which is related to the number of markers to be used; and (4) the relationship between molecular genetic distance and distance based on morphological descriptors. Elements concerning (3) have already been developed by Bar-Hen and Charcosset (1995). The aim of the study presented here was to bring together the elements of those four points based on experimental results in maize. For this purpose, 145 maize inbred lines, representative of material released in France, were described using both restriction fragment length polymorphism (RFLP) markers and a set of morphological traits used in current distinctness studies.

Materials and methods

Germplasm under study

We used 145 maize inbred lines for this study. These inbred lines were obtained in various breeding programs run by private companies, as well as public institutes, and they are adapted to the various climatic conditions found in France. They were chosen to be representative of the different groups of maize germplasm used in Europe. For reasons of confidentiality, the inbred lines were coded and the precise relatedness between them is unknown.

RFLP analyses

Out of the 250 genomic clones kindly provided by the Brookhaven National Laboratory (BNL probes) and the University of Missouri (UMC probes) a subset of 100 was chosen for DNA profiling (Table 1). The two criteria used to select the subset of 100 probes were the quality of the hybridization signal and the genome coverage. The RFLP protocol described in Murigneux et al. (1993) was used except for a few modifications. DNA was extracted from 15 young plants and then digested with three restriction enzymes (*Eco*RI, *Hin*dIII, *Eco*RV). Then, 5 μg restricted and phenol : chloroform-purified DNA was loaded per lane. One nanogram of a 1.1-kb human *Alu* fragment was added to each DNA sample prior to electrophoresis. In three of the lanes per gel, a molecular-weight (MW) marker consisting of ten bands of lambda phage ranging from 24.8 to 0.8 kb in size was loaded. The MW fragments were cloned and amplified separately. As each fragment was added at an equivalent of 1 ng per lane, an equal intensity was obtained for each band. The autoradiograms were scanned and automatically scored with a bioprofil software (Vilber Lourmat) which provided the molecular weight of the bands. For each of the 145 inbred lines, the analysis was performed on each of the (3×100) enzyme $*$ probe combinations (EPC). In addition, 5 inbred lines were repeated and therefore were studied twice in a double-blind protocol, leading to a total of $(145 + 5) \times 3 \times 100 = 45\,000$ different profiles. Among the 300 EPCs, only 222 provided interpretable profiles (Table 1). The position of the monolocus probes on the maize genetic map (Table 1) was obtained from several maize populations (BIOCEM, unpublished).

Comparison of the information provided by the markers

To compare the information given by the different enzyme $*$ probe combinations we computed Nei diversity index (Nei and Roychoudhury 1973) for each EPC as:

$$h_i = 1 - \sum_j p_{ij}^2 \tag{1}$$

where p_{ij} is the frequency of the jth profile with the ith EPC.

For monolocus probes and inbred lines, profiles were considered to be alleles. For each probe, the information revealed by the three enzymes can be described in a three-way contingency table by computing the number of inbred lines which have profile i with the enzyme *Eco*RI, profile j with the enzyme *Hin*dIII, and profile k with the enzyme *Eco*RV. Alternatively, the same information can be described by three different two-way contingency tables for each probe. To quantify the redundancy of the information, there are many possible ways to compute a coefficient of association for this kind of table of contingency. Since the number of inbred lines is constant (145) and the number of profiles is variable (and therefore also the number of rows and columns in the contingency table), we computed Cramer's V coefficient of association (see Bishop 1975):

$$V = \sqrt{(\chi^2_{(r-1)(c-1)}/n) \min(r-1, c-1)} \tag{2}$$

where $\chi^2_{(r-1)(c-1)}$ is the value of the classical chi-square statistics of independence for $r * c$ contingency table based on n observations.

To qualify the specific information provided by each probe, we can consider each combination of profiles revealed by the three enzymes for the same probe as one allele. Hence, we defined the effective number of alleles n_E as the number of non-empty cells of the three-way $r * c * l$ contingency table given by the three enzymes. The minimum value of this quantity is equal to $n_R = Max(r, c, l)$ and corresponds to the case of maximum redundancy with complete association between profiles revealed by different enzymes. The maximum value is equal to the product $r * c * l$ and corresponds to the case of independent enzymes for large sample sizes. The ratio n_E/n_R gives an indication of the information attained using the combination of the three enzymes relative to that obtained by choosing the most informative enzyme among the three.

Computation of marker distances

There are several ways of computing a distance index between two inbred lines. It mainly depends on whether the bands are interpreted as alleles or not. In the ideal situation, each RFLP band is associated with one allele, and all distance indices are equivalent. In reality, a given profile can be characterized by the variation of several bands, and the distance index can be computed either on band information or on profile information.

The relevant parameter to estimate here is the percentage d_{XY} of loci which differ between two inbred lines X and Y. It is directly related to the coancestry coefficient between the inbred lines, as defined by Malecot (1948). For that purpose, two different kind of distance indices were computed. The first one is Nei's genetic distance (Nei and Li 1979) computed on band information:

$$Nei = 1 - \frac{2N_{XY}}{(N_X + N_Y)} \tag{3}$$

where N_X (respectively N_Y) is the number of bands found in inbred line X (respectively Y), and N_{XY} is the number of bands shared by the inbred lines X and Y. The second kind of distance index was Rogers distance (Rogers 1972) computed on monolocus EPCs on the basis of allelic information,

$$MRD = \frac{1}{2L_m} \sum_{i=1}^{L_m} \sum_{j=1}^{n_i} (p_{ij}^X - p_{ij}^Y)^2 \tag{4}$$

where L_m is the number of monolocus EPCs involved in the study, n_i is the number of alleles for the ith EPC, and p_{ij}^X (respectively p_{ij}^Y) is the frequency of the jth allele for inbred line X (respectively inbred line Y). Note that with inbred lines, Rogers distance reduces to the simple matching coefficient (Gower 1985). Finally, a synthetic distance index was proposed in order to combine information on both monolocus and multilocus EPCs as

$$DI = \frac{L_m}{L_m + L_M} MRD + \frac{L_M}{L_m + L_M} Nei \tag{5}$$

where Nei is computed from the band information carried by L_M multilocus EPC.

Supposing that EPCs are sampled at random over the genome, the expectation of MRD is d_{XY}, and its sampling variance is equal to $d_{XY}(1 - d_{XY})/L_m$, which can be estimated by replacing d_{XY} by MRD (Bar-Hen and Charcosset 1995). Without any idea about the distribution of the number of bands within 1 inbred line, the sampling variance of Nei cannot be computed exactly. A rough approximation can be obtained by $Nei(1 - Nei)/\bar{N}$, where $\bar{N}$ is the average number of bands per inbred line.

Morphological description and distance computation

Morphological data for ten quantitative traits (Table 2) were collected at three locations in France (La Minière near Paris, Le

Table 1 Description of the 222 interpretable enzyme * probe combinations

Probe		Ch[a]	Map[b] (cM)	Enzyme[c]	N[d]	Quality	n_i[e]	Hi[f]	nE[g]
BNL5-62		1		E	?	B	6	0.573	
BNL5-62		1		H	1	B	5	0.558	
BNL5-62	x	1	0	V	1	B	9	0.769	23
UMC157	x	1	29	E	1	B	8	0.589	16
UMC157		1		H	?	C	8	0.599	
UMC157		1		V	1	B	5	0.624	
UMC11	x	1	58	H	1	B	11	0.759	13
UMC11		1		V	1	B	4	0.658	
BNL12-06		1		E	1	B	8	0.726	
BNL12-06	x	1	76	H	1	B	9	0.726	26
BNL12-06		1		V	1	B	7	0.669	
UMC67		1		H	1	B	4	0.303	
UMC67	x	1	91	V	1	B	6	0.655	9
UMC128		1		E	1	C	8	0.54	
UMC128		1		H	1	C	8	0.636	
UMC128	x	1	125	V	1	B	7	0.474	17
UMC83	x	1	134	H	1	C	7	0.698	11
UMC83		1		V	1	C	5	0.377	
UMC107		1		E	1	B	3	0.449	
UMC107	x	1	148	H	1	A	2	0.44	5
UMC107		1		V	1	B	4	0.502	
UMC106	x	1	158	E	1	B	5	0.563	9
UMC106		1		H	1	C	5	0.517	
UMC106		1		V	1	C	2	0.5	
BNL7-25		1		E	1	C	4	0.498	
BNL7-25	x	1	160	H	1	B	4	0.369	25
BNL7-25		1		V	?	C	11	0.817	
BNL8-29	x	1	166	E	1	A	5	0.375	12
BNL8-29		1		H	?	C	9	0.7	
UMC61		2		E	1	C	4	0.614	
UMC61	x	2	0	H	1	C	7	0.665	9
UMC34		2		E	1	B	3	0.026	
UMC34		2		H	1	C	9	0.679	
UMC34	x	2	17	V	1	B	5	0.439	13
BNL12-09	x	2	20	H	1	A	4	0.525	10
BNL12-09		2		V	1	B	8	0.763	
UMC131		2		E	?	B	4	0.569	
UMC131	x	2	35	H	1	A	5	0.473	8
UMC131		2		V	1	B	2	0.169	
UMC55		2		E	?	A	8	0.735	
UMC55		2		H	1	B	5	0.163	
UMC55	x	2	43	V	1	A	4	0.478	15
UMC136		2		H	1	B	5	0.699	
UMC136	x	2	49	V	1	A	5	0.555	8
UMC4	x	2	62	E	1	A	5	0.555	10
UMC4		2		H	1	B	5	0.676	
UMC4		2		V	1	B	8	0.795	
UMC122	x	2	65	H	1	A	5	0.659	6
UMC122		2		V	1	B	2	0.026	
UMC139	x	2	77	H	1	C	6	0.459	11
UMC139		2		V	1	C	4	0.614	
UMC32	x	3	0	E	1	B	6	0.587	14
UMC32		3		H	1	B	6	0.649	
UMC32		3		V	?	A	10	0.646	
UMC121	x	3	16	E	1	A	9	0.732	22
UMC121		3		H	1	B	5	0.599	
UMC121		3		V	1	C	9	0.768	
UMC10		3		E	1	B	9	0.694	
UMC10	x	3	54	H	1	A	3	0.482	17
UMC10		3		V	1	C	4	0.728	
UMC50		3		E	?	B	10	0.684	
UMC50	x	3	59	V	1	B	11	0.69	16
UMC102	x	3	61	E	1	A	6	0.561	14
UMC102		3		H	1	A	5	0.209	
UMC102		3		V	1	B	13	0.657	
BNL6-06		3		E	1	B	6	0.722	
BNL6-06	x	3	65	H	1	A	10	0.734	20
BNL6-06		3		V	1	B	9	0.722	
BNL5-37	x	3	78	E	1	B	10	0.759	17
BNL5-37		3		H	1	B	8	0.485	
BNL5-37		3		V	1	B	8	0.36	
UMC60	x	3	94	E	1	A	6	0.612	13
UMC60		3		H	1	A	3	0.285	
UMC60		3		V	1	B	5	0.53	
UMC63		3		E	1	C	5	0.614	
UMC63	x	3	152	H	1	A	6	0.687	13
UMC63		3		V	1	C	6	0.583	
UMC31		4		H	1	B	2	0.147	
UMC31	x	4	0	V	1	B	5	0.487	6
ADH2	x	4	10	H	1	A	8	0.651	18
ADH2		4		V	?	A	15	0.784	
UMC66		4		E	?	B	8	0.734	
UMC66	x	4	43	H	1	C	6	0.522	13
UMC19		4		E	?	C	9	0.767	
UMC19	x	4	49	H	1	A	6	0.409	23
UMC19		4		V	1	C	7	0.536	
BNL7-65	x	4	66	E	1	B	12	0.763	20
BNL7-65		4		H	1	B	5	0.635	
BNL7-65		4		V	1	C	7	0.399	
UMC15		4		E	1	B	6	0.429	
UMC15	x	4	74	H	1	B	7	0.464	10
UMC15		4		V	1	C	5	0.401	
BNL6-25	x	5	9	H	1	B	9	0.481	12
BNL6-25		5		V	1	B	5	0.391	
UMC90	x	5	30	E	1	A	4	0.604	12
UMC90		5		H	1	B	5	0.708	
UMC90		5		V	?	B	10	0.737	
UMC27		5		E	?	C	8	0.791	
UMC27	x	5	51	H	1	B	4	0.731	16
UMC27		5		V	1	C	5	0.625	
BNL7-56	x	5	55	H	1	B	4	0.364	4
BNL6-22		5		E	1	B	7	0.606	
BNL6-22	x	5	73	H	1	A	2	0.039	10
BNL6-22		5		V	1	B	3	0.148	
BNL10-12		5		E	?	C	6	0.745	
BNL10-12	x	5	96	H	1	A	4	0.263	12
BNL10-12		5		V	1	C	7	0.626	
BNL7-71	x	5	99	E	1	B	6	0.629	14
BNL7-71		5		H	1	A	6	0.531	
BNL7-71		5		V	1	B	4	0.472	

Table 1 continued

Probe		Ch[a]	Map[b] (cM)	Enzyme[c]	N[d]	Quality	n_i[e]	Hi[f]	nE[g]
UMC51	x	5	104	E	1	B	13	0.718	19
UMC51		5		H	1	C	5	0.538	
UMC126		5		E	?	B	6	0.528	
UMC126	x	5	104	H	1	B	10	0.676	17
UMC126		5		V	1	B	6	0.599	
UMC104		5		H	1	B	7	0.567	
UMC104	x	5	169	V	1	B	9	0.646	15
UMC85		6		H	1	B	4	0.617	
UMC85	x	6	0	V	1	A	7	0.678	9
UMC59	x	6	8	E	1	B	4	0.227	9
UMC59		6		H	1	C	5	0.696	
UMC65		6		E	1	B	4	0.583	
UMC65		6		H	?	B	7	0.607	
UMC65	x	6	31	V	1	B	7	0.674	13
UMC21	x	6	42	E	1	A	8	0.745	9
UMC21		6		H	1	A	5	0.59	
BNL3-03		6		E	?	B	8	0.759	
BNL3-03	x	6	49	H	1	B	8	0.548	12
UMC38	x	6	67	H	1	B	7	0.673	7
UMC132		6		E	1	C	2	0.358	
UMC132		6		H	1	B	2	0.358	
UMC132	x	6	91	V	1	A	5	0.612	6
UMC62		6		E	1	B	3	0.349	
UMC62	x	6	106	H	1	B	4	0.172	6
UMC62		6		V	1	B	4	0.369	
UMC167		6		E	1	B	3	0.506	
UMC167	x	6	117	H	1	B	5	0.607	7
UMC116		7		E	?	C	19	0.856	
UMC116		7		H	1	B	6	0.66	
UMC116	x	7	11	V	1	B	8	0.78	26
UMC110	x	7	25	E	1	B	17	0.913	23
UMC110		7		H	1	C	8	0.77	
BNL6-27	x	7	30	V	1	A	2	0.498	2
BNL5-21		7		E	2	C	19	0.988	
BNL5-21		7		H	2	C	12	0.826	
BNL5-21	x	7	46	V	1	B	3	0.453	45
BNL7-61		7		H	1	B	4	0.173	
BNL7-61	x	7	52	V	1	B	7	0.74	9
UMC151	x	7	64	H	1	B	9	0.738	14
UMC151		7		V	1	C	8	0.367	
BNL16-06		7		H	?	B	16	0.788	
BNL16-06	x	7	70	V	1	C	6	0.51	20
UMC35		7		E	1	B	2	0.241	
UMC35	x	7	87	H	1	A	6	0.519	28
UMC35		7		V	?	C	20	0.866	
UMC168		7		E	?	A	6	0.679	
UMC168	x	7	93	V	1	B	5	0.547	10
BNL9-11		8		H	?	B	6	0.404	
BNL9-11	x	8	19	V	1	A	6	0.688	10
UMC103		8		E	1	C	8	0.685	
UMC103	x	8	31	H	1	A	7	0.659	17
UMC103		8		V	1	C	10	0.715	
BNL9-08	x	8	50	H	1	B	4	0.474	17
BNL9-08		8		V	?	B	12	0.783	
UMC89		8		H	1	C	5	0.487	
UMC89	x	8	75	V	1	A	6	0.593	8
UMC48	x	8	86	H	1	B	24	0.923	43
UMC48		8		V	1	C	11	0.794	
UMC109		9		E	?	C	4	0.304	
UMC109	x	9	0	H	1	A	3	0.039	8
UMC109		9		V	1	B	7	0.519	
UMC113	x	9	15	E	1	B	10	0.719	10
UMC114		9		H	1	C	6	0.596	
UMC114		9		V	1	C	6	0.591	
UMC140	x	9	80	H	1	B	4	0.585	4
UMC140		9		V	1	B	3	0.457	
BNL7-57		9		E	?	C	11	0.845	
BNL7-57		9		H	1	C	16	0.796	
BNL7-57	x	9	113	V	1	B	9	0.693	28
BNL14-28		9		H	1	A	6	0.6	
BNL14-28	x	9	116	V	1	A	11	0.669	13
BNL5-09		9		E	1	A	5	0.536	
BNL5-09	x	9	120	H	1	A	9	0.732	12
BNL5-09		9		V	1	B	8	0.705	
BNL3-04	x	10	0	H	1	B	6	0.463	9
BNL3-04		10		V	?	C	4	0.467	
UMC155		10		H	1	B	5	0.566	
UMC155	x	10	57	V	1	A	6	0.642	7
UMC146		10		E	1	B	5	0.563	
UMC146	x	10	63	H	1	A	4	0.149	6
UMC57	x	10	88	E	1	B	5	0.491	9
UMC57		10		H	1	B	4	0.621	
R1		10		E	1	B	5	0.395	
R1	x	10	90	H	1	B	7	0.679	10
UMC86	x			H	?	C	2	0.43	
BNL13-05	x			H	2	C	16	0.845	
BNL13-05				V	?	C	6	0.451	
BNL2-369	x			H	?	B	8	0.749	
BNL5-10				E	?	B	8	0.703	
BNL5-10	x			H	2	A	8	0.692	
BNL5-10				V	2	A	5	0.66	
BNL5-71				H	?	A	7	0.599	
BNL5-71	x			V	?	B	8	0.636	
BNL7-20	x			E	?	B	5	0.511	
BNL7-20				V	?	C	8	0.37	
BNL8-33	x			E	?	C	10	0.694	
BNL9-44				E	?	C	5	0.749	
BNL9-44	x			H	?	A	6	0.725	
BNL9-44				V	?	B	4	0.591	
UMC130	x			V	?	B	6	0.783	
UMC16				E	0	B	16	0.763	
UMC16	x			H	0	A	14	0.677	
UMC16				V	0	C	25	0.92	
UMC3	x			E	2	B	13	0.642	
UMC3				V	2	C	14	0.884	
UMC44	x			H	2	B	16	0.841	
UMC44				V	2	C	12	0.866	

Table 1 continued

Probe		Ch[a]	Map[b] (cM)	Enzyme[c]	N[d]	Quality	n_i[e]	Hi[f]	nE[g]
UMC47	x			E	?	A	3	0.17	
UMC81	x			H	?	A	7	0.659	
UMC81				V	?	A	4	0.677	
UMC98	x			H	?	C	3	0.498	

[a] Chromosome
[b] Position of the probe on the maize genetic map. ≪x≫ indicates the EPCs chosen for the computation of MRD or Nei78
[c] E, *Eco*RI; H, *Hin*dIII; V, *Eco*RV
[d] Number of loci involved. ?, multilocus EPC
[e] Number of alleles detected
[f] Nei diversity index
[g] Effective number of alleles. nE has been placed arbitrarily beside a spesific EPC but has a bearing on all the enzymes for each probe

Magneraud in the center-west and Saint Martin de Hinx in the south-west) during the years 1989, 1990, 1991, and 1992. Depending on their earliness, 75 inbred lines were evaluated for at least 2 years in at least two convenient locations. The remaining 70 inbred lines were evaluated in La Minière and another location, depending on earliness, in 1992. Each location was planted with two replications in a block design. Within each block, a single plot of each inbred was grown with 20 individuals spaced in a row 4.75 m long with 80 cm between the rows; traits were measured individually on the central 10 plants of each row and then averaged. The experimental design comprised a total of 1376 elementary plots. An analysis of variance was performed on each trait with three main effects: location, year, and block effect, and the location * year interaction. The fraction of the variation due to controlled environmental effects was estimated by R^2, the coefficient of determination of the model (Table 2). The residuals of this model provide an estimate of the phenotypic effect of each inbred line (Bar-Hen et al. 1995). To estimate the fraction R_m^2 of phenotypic variance accounted for by the inbred line effect, we performed a second analysis of variance on those residuals, weighted by the number of replications of each inbred line within the experimental design (Bar-Hen et al. 1995). Mahalanobis distance (Mahalanobis 1936) was computed for each couple of inbred lines based on the residuals of the first model, i.e., on the phenotypic effect of the inbred lines.

Experimental results

Description of polymorphism at marker loci

A good quality of markers is fundamental to obtaining reproducible results and reliable distances between varieties. Depending on the quality of the result, we graded the enzyme * probe combinations as A (very good quality) to *D* (non-interpretable). Table 3 gives the repartition of the probes with respect to the quality and the enzyme used. Out of 300 EPCs, a total of 222 EPCs corresponding to 95 probes were polymorphic and interpretable (i.e., quality C at least). The average number of levels of migration (band levels) for the 222 interpretable markers was 4.96. It varied from 4.43 for markers of quality A to 5.36 for markers of quality C, which corresponded to a total of 1098 bands.

To evaluate the precision of the method, we repeated the study with 5 inbred lines, and therefore these were studied twice in a double-blind protocol. The number of differences over the 222 EPCs ranged from zero for 1 replicated inbred line to seven for another inbred line, with 2 replicated lines exhibiting one difference and the last one exhibiting three differences. Six of these differences were observed with markers of quality C. The differences appeared on EPCs corresponding to different probes, except for the inbred line with seven differences, for which 2 probes revealed differences with two enzymes. These discrepancies between replications may come either from an experimental error or from a residual heterozygosity of the inbred lines.

The discrepancy may be considered as a residual error attached to the use of molecular markers, occurring for each EPC with probability p_e. This probability may be estimated from our five replications. It is equal to 0.0108 among the 222 EPCs of quality A to C and to 0.0049 among the 163 EPCs of quality A or B. Hence, if

Table 2 Analysis of the quantitative traits data

Name of the trait	Mean	Variance[a]	R^2[b]	R_m^2[c]
Ear: length (mm)	149.34	331.28	0.08	0.85
Ear: diameter of ear (mm)	41.22	12.20	0.12	0.86
Ear: diameter of cob (mm)	26.77	5.45	0.40	0.76
Ear: Number of rows of seeds	14.06	3.09	0.07	0.88
Plant: Height of ear (cm)	62.14	154.95	0.28	0.82
Plant: Total plant length (cm)	148.90	438.61	0.21	0.87
Leaf: Width of blade (mm)	90.97	87.94	0.24	0.75
Tassel: Length of main axis above lowest side branch (cm)	29.57	13.83	0.07	0.86
Tassel: Length of main axis above highest side branch (cm)	20.76	12.10	0.04	0.85
Tassel: Date of male flowering (days since January 1st)	207.43	17.86	0.74	0.82

[a] Phenotypic variance of inbred lines
[b] Part of the variation due to controlled environmental effects of the model
[c] Part of the phenotypic variance among inbred lines accounted for by the inbred line effect

Table 3 Quality of the 300 enzyme * probe combinations

Quality	*Eco*RI	*Hin*dIII	*Eco*RV	Total
A	11	25	13	49
B	35	42	37	114
C	15	19	25	59
D	39	14	25	78
Total	100	100	100	300

d_{XY} is the real genetic distance between inbred lines X and Y, there are two sources of variation for the estimation of d_{XY} from molecular data. The first one is the sampling of the markers throughout the genome, with associated sampling variance $d_{XY}(1-d_{XY})/L$, where L is the number of markers involved. The second source of variation is the residual error, with associated error variance $p_e(1-p_e)/L$. The reliability r^2 of the results can be computed as the ratio of the sampling variance to the total variance,

$$r^2 = 1 - \frac{p_e(1-p_e)}{(p_e(1-p_e)+d_{XY}(1-d_{XY}))} \quad (6)$$

It ranges from 0.816 with $d_{XY}=0.05$ to 0.959 with $d_{XY}=0.5$ among the 222 EPCs. Eliminating the 59 EPCs of quality C enhances the reliability, then ranging from 0.930 to 0.986.

The 222 EPCs represent a total of 1546 profiles. (442 for *Eco*RI, 570 for *Hin*dIII and 534 *Eco*RV). Therefore, there is an average of seven profiles per interpretable enzyme * probe combination. A total of 421 specific profiles are present in only 1 line; 106 inbred lines have at least 1 specific profile. The case of line 50 is also of interest: its most characteristic profile is also present in 13 other inbred lines. This line could be either a founder line, or it may contain a combination of common profiles.

Markers were also classified according to the number of loci involved in the variation of the patterns (Table 1). One hundred and sixty-seven EPCs are monolocus and correspond to 80 different loci, while 9 EPCs involve 2 loci. For the remaining 46, it was not possible to draw a conclusion about the number of loci involved. The average map distance between 2 neighbour loci is 13.4 cM for a total map length of 1143 cM. Nei's diversity index was computed for each EPC. It is equal to the probability of having two different alleles in 2 different lines drawn at random among the 145 and ranges from 0.03 to 0.92. However, only 3 EPCs, corresponding to probes UMC122, BNL6-22 and UMC109, have a really low Nei's diversity index with one enzyme. The average value for Nei's diversity index over the enzymes for each monolocus probe is represented as a function of the position of the markers on the genetic map (Fig. 1). It ranges from 0.26 to 0.86, with an average value of 0.55. This indicates that, except for some regions on chromosomes 5 and 6, the average relatedness among the material released in France is relatively low. These results are comparable to those obtained on similar material by Dubreuil et al. (1996).

Redundancy of the information

Cramer's V coefficient of association between *Eco*RI and *Hin*dIII was computed for 54 probes, the association between *Eco*RI and *Eco*RV was computed for 45 probes, and the association between *Hin*dIII and *Eco*RV for 68 probes. The theoretical range of V is between 0 and 1. The associations found were always high and positive. For the three pairs of enzymes, the median of the V is always between 0.7 and 0.8. The discrepancy around these values is comparable for the three pairs of enzymes. This suggests that, in general, the use of a single enzyme per probe can be recommended, since a second enzyme brings only a little additional information.

However, an alternative approach could be to combine the information of the different enzymes for each probe by considering profile combinations as allelic forms. For each of the 80 monolocus probes, the corresponding effective number of alleles, n_E, was computed (Table 1 and Fig. 1). The ratio n_E/n_R of the effective number of alleles on the number of alleles obtained by choosing the most informative enzyme for each probe was also computed. It varies between 1 and 2.89, with an average value of 1.57. This ratio is higher than 2 for about one-third of the probes, which means that the number of alleles can be increased without redundancy by a factor two by taking into account the information given by different enzymes with the same probe.

Four distance indices were computed. Primarily, *Nei* genetic distance was computed over the 1091 bands obtained from the 222 EPC. Secondly, one enzyme per probe was chosen for the 80 monolocus probes on the basis of two criteria: the EPC should be of quality A or B and, for equal quality, should reveal a maximum number of alleles. The enzyme chosen for each probe is indicated (Table 1). The distance index MRD was computed on this subset. Thirdly, in order to use all the information, we constituted a second data set with the 80 monolocus probes. For each probe, the information on enzymes of quality A or B was combined into new alleles. A distance index MRD' was computed as MRD on this new subset, which involved 136 EPCs and 80 different probes. Finally, multilocus EPCs, for which allelic interpretation was not possible, were incorporated into a synthetic distance. To avoid redundancy, one enzyme of quality A or B was chosen for each multilocus probe, and Nei_{78} genetic distance was computed on the basis of band information. There were 78 bands corresponding to 15 different probes. The synthetic distance index DI was computed as

$$DI = \frac{(80MRD' + 15Nei_{78})}{95} \quad (7)$$

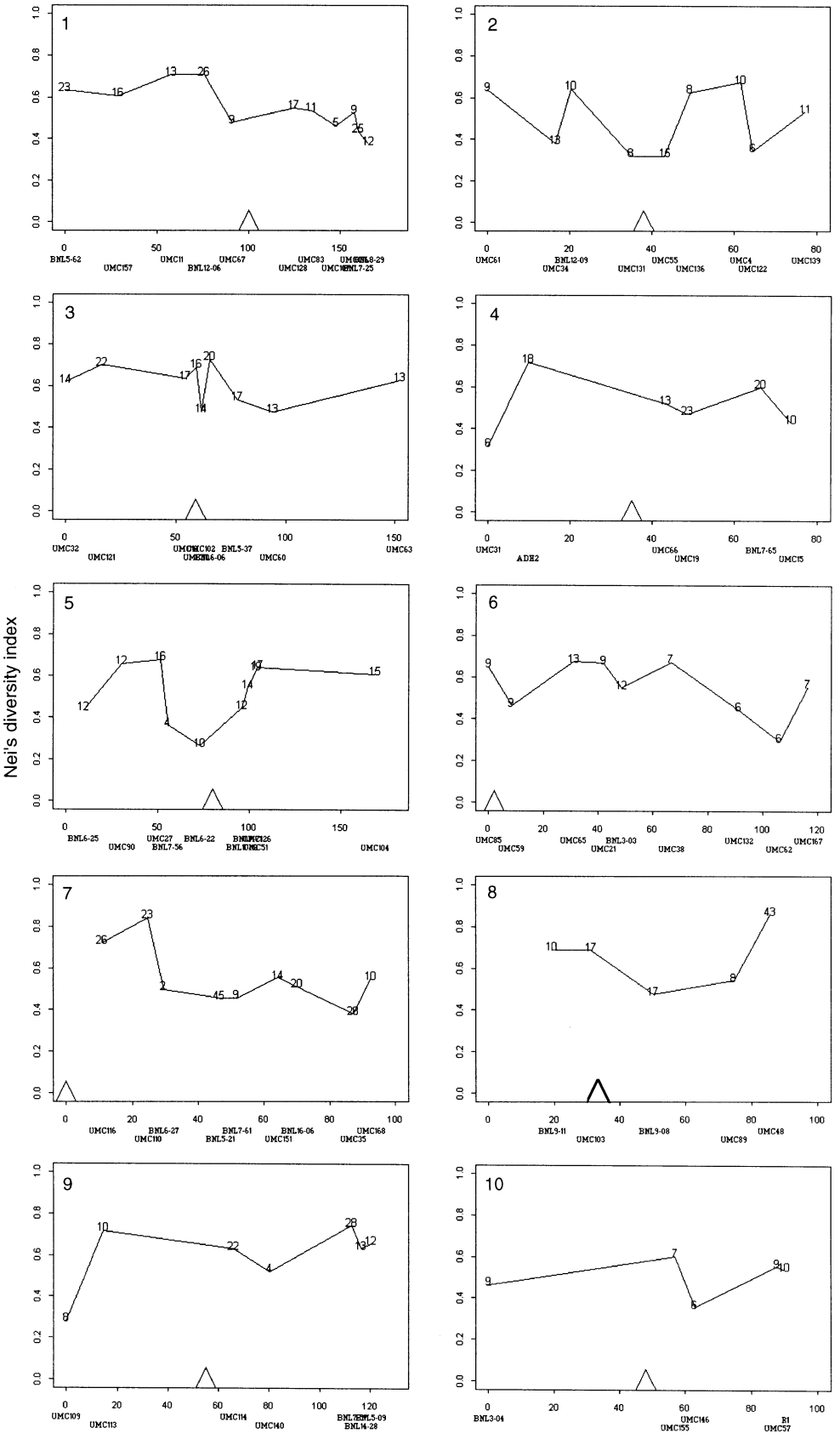

Fig. 1 Relationship between Nei's diversity index and the position of the monolocus markers on the genetic map. The *numbers above* each *point* are the effective number of alleles for each probe. *Triangles* indicate the approximate position of the centromere. The *numbers* in the *upper-left corner* of each *box* represents the number of the maize chromosome

Table 4 Elementary statistics of the molecular distance indices[a]

	Nei	*MRD*	*MRD′*	*DI*
Minimum	0.004	0.000	0.000	0.000
Maximum	0.702	0.837	0.925	0.851
Mean	0.523	0.577	0.642	0.612
SD	0.084	0.100	0.107	0.100
Maximum sampling SD	0.031	0.056	0.056	0.050
Mean sampling SD	0.031	0.054	0.052	0.047

[a] See text for the computation of the sampling variances

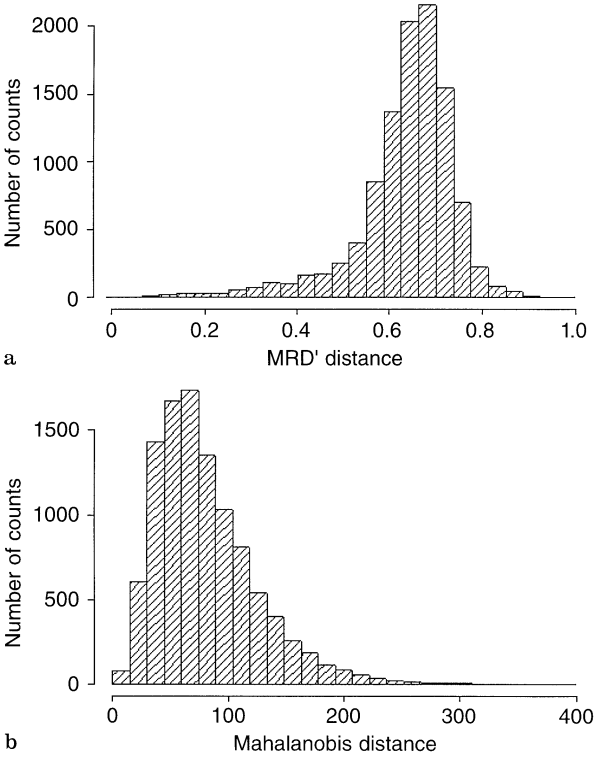

Fig. 2a, b Distribution of the molecular (*MRD′*) (**a**) and morphological (Mahalanobis) (**b**) genetic distances among the couples of inbred lines. Scaling based on class frequencies

Evaluation of marker distance

Four distance indices were computed for each of the $145 \times 144/2 = 10\,440$ couples of inbred lines (Table 4). There is 1 couple of inbred lines which differs only for 1 EPC (UMC34 with *Hin*dIII) of quality C. Hence, the minimum distance is zero for both *MRD*, *MRD′*, and *DI*, for which the markers of quality C were discarded. The maximum distance is obtained with *MRD′*, which combines the information of several enzymes per probe. The average distance varies between 0.523 with *Nei* to 0.64 with *MRD′*. Note that the average value of 0.58 for *MRD* is comparable to the average Nei's diversity index, the two parameters measuring the same quantity. The average distance is lower with *Nei* than with the other distance indices. As a matter of fact, it was computed on band information over the 222 EPCs of the study. The lower genetic distances observed can be due to redundancy of band information for multiple band profiles, i.e., to different alleles sharing the same band.

The distribution of *MRD′* among the 10 440 couples of inbred lines is representative of the distribution of all the distance indices (Fig. 2a). The skewness towards low genetic distances indicates the existence of some couples that represent highly related inbred lines.

The widest ranges of distance indices were obtained with the three distance indices *MRD*, *MRD′* and *DI* (Fig. 3). As *MRD′* considers more alleles at a given locus, it uses more information than *MRD*, and *MRD′* appears to be consistently more discriminant than *MRD*. On the contrary, the synthetic index *DI* is not always more discriminating than *MRD* and can even be lower than *MRD*. The discrepancy between the distance indices increases with the genetic distance between inbred lines (Fig. 3). A comparison of Fig. 3a and b illustrates the effects of redundancy. As the genetic distance increases, both the relatedness between the inbred lines decreases and the correlation between independent profiles or independent bands decreases, so that the difference between distance indices increases (Fig. 3a). On the contrary, non-independent information introduces an upper limit for the genetic distance, observed with *DI* (Fig. 3b). In this case, including the band information carried by multilocus EPC in a synthetic index, *DI* clearly introduces redundancy. Band treatment of the information definitely appears as being less discriminating than allelic treatment, especially if 'synthetic alleles' are defined by combining the information revealed by different enzymes for the same probe.

Sampling variances were computed for each distance index (Table 4). Note that the variances of *MRD* and *MRD′* among the 10 440 couples of inbred lines are about four times greater than the corresponding average sampling variance. This indicates the diversity of relatedness between the inbred lines. As the number of loci is relatively high, the binomial distribution of *MRD* can be approximated by the normal distribution. Then, the approximate 5% level confidence interval is given by

$$MRD \pm 1.96 \sqrt{\frac{MRD(1-MRD)}{80}} \tag{8}$$

The sampling variance has been reduced here by selecting the markers based on their position on the genetic map. In this case, the value of the sampling variance would also depend on the relatedness between inbred lines. Hence, sampling variances given here only constitute upper bound values (Bar-Hen and Charcosset 1995).

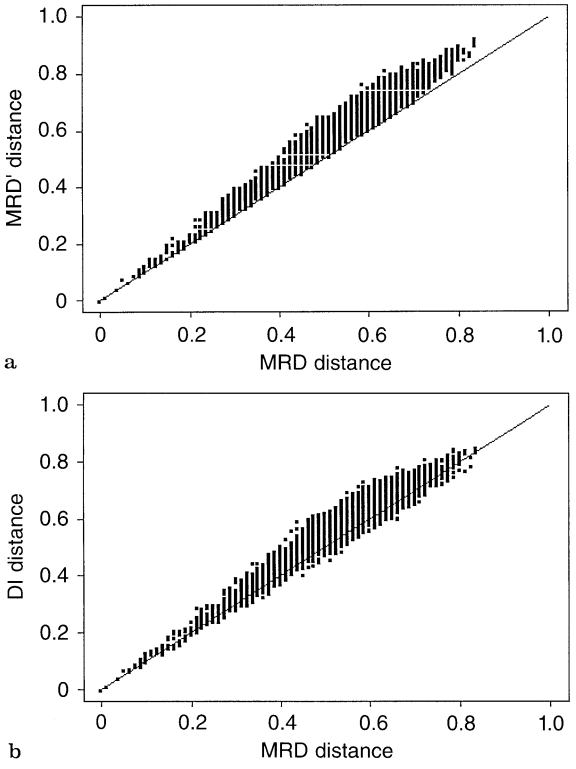

Fig. 3 Relationship between molecular genetic distances. *MRD* is based on 80 monolocus probes with 1 enzyme per probe; *MRD′* uses the information of 136 EPC from 80 monolocus probes without redundancy and *DI* is a synthetic index combining both monolocus and multilocus probes (see text for details). The straight line is the $x = y$ line

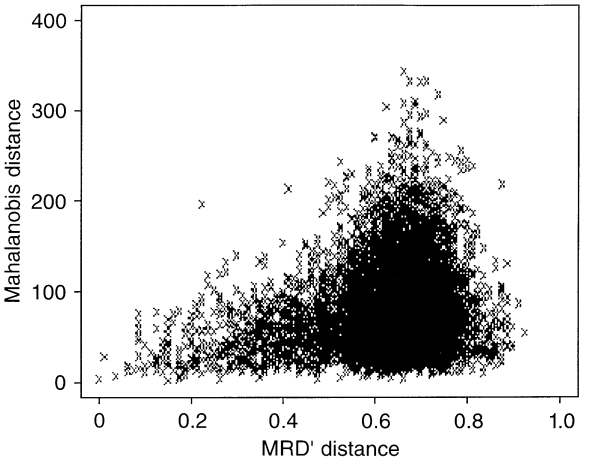

Fig. 4 Relationship between molecular (*MRD′*) and morphological (Mahalanobis) genetic distances

Evaluation of morphological distance and its relationship to marker distance

Analysis of variance was performed on each of the ten morphological traits studied (Table 2). Most of them appeared to be midly affected by year and location effects, except for diameter of the cob ($R^2 = 0.40$) and date of male flowering ($R^2 = 0.74$). The inbred line effect was always important and explained a fraction of the phenotypic variance, ranging from 0.74 for width of the blade to 0.88 for number of rows of seeds (Table 2). The Mahalanobis distance based on those ten morphological traits was also computed for each of the 10 440 couples of inbred lines. It ranges from 3 to 344 (Fig. 2b) with an average value of 80. It's distribution is skewed, due to a few couples of very different inbred lines. The patterns of variation of morphological and molecular distances are quite different, even though they both cover a wide range of distance values (Fig. 2a and b).

Figure 4 gives the relationship between marker distance and morphological distance. It is clearly not linear but displays a "triangular" shape. Low marker distances are systematically associated with low morphological distances. On the other hand, high marker distances are associated either to high or low morphological distances. Thus, marker divergence behaves as a limiting factor of morphological divergence.

Discussion and conclusion

Optimization of RFLP experiments for diversity analysis and distinctness studies

The quality required for RFLP data depends on the objectives of the investigation. With respect to diversity analysis and distinctness studies, the objective is to perform routine studies on large data sets, with readily interpretable molecular data. Hence, both a repeatable standard protocol and good quality probes are required.

The final quality of the autoradiograms is dependent on the quality and quantity of both target and probe DNA, and on the control of all the parameters during electrophoresis and hybridization. In this experiment, the use of the *Alu* band allowed us to check for potential problems during electrophoresis. The use of a reliable molecular-weight marker is necessary for the automatic estimation of the molecular weight of a specific band. This system also enables the comparison of results produced on different autoradiograms. As for the quality of the probes, the most important criterium is to distinguish easily between 2 or more different bands. For example, some probes revealed as many as 28 different patterns involving 12 different bands. While being very informative, the differences between the electrophoretic mobility of those bands become very small and the risks of misassignment of the bands increase. Therefore, the selection of probes providing not only enough but also reliable information is required. Before selection of such quality probes can be achieved for the whole genome, it is suggested that very stringent hybridization conditions be used. Even with

a good protocol, we have shown that it is impossible to obtain 100% reliable results. By analogy with the definition of heritability, we have proposed a reliability parameter. This parameter would depend on both the protocol, which may be improved, and on the residual heterozygosity, which is an inherent characteristic of the material.

Choice of the probes and restriction enzymes

For the molecular profiles of the varieties to be entered into a database, a standard set of markers needs to be defined. After the elimination of probes exhibiting bands that cannot be discriminated from one another, several criteria may be used to choose a reasonable set of probes given the cost of an experiment. Primarily, the number of markers will depend on the precision to be reached for the estimation of genetic distance. Each marker can be considered as one sample of the genome and, according to Eq. 8, the precision of the estimate increases with the square root of the number of markers. It increases very quickly between 0 and 100, and increases somewhat slower above 100 markers. If the molecular genetic distance is to be compared to a minimum genetic distance useful to the question of intellectual property protection, the number of markers can also be determined by examining the two levels of the test: the probabilities of declaring the genetic distance below and above the minimum distance given reality. Those probabilities depend on the sampling variance of the estimate, but also on the variance of errors coming from experimental error or residual heterozygosity. The latter can be determined from replications of the inbred lines. In our experiment, the variance of errors was shown to be low, especially when probes of good quality were chosen.

The second criteria is the repartition of the markers on the genetic map of the species. As a matter of fact, it can be dangerous for plant breeding to rely on a restricted part of the genome for any germplasm evaluation. Moreover, selection of a set of markers based on their position on the genetic map by practising some kind of stratified sampling should reduce the sampling variance of genetic distance (Cochran 1977), especially for related inbred lines. We are currently investigating that point in order to find a method of assessing the precision of estimates for related material. In the ideal situation, markers should be evenly spaced on each chromosome, but this is hardly ever attained in actual practice. The choice between two adjacent markers can be made on the basis of their discriminant power in a reference collection of inbred lines (Fig. 1).

Finally, genetic determinism of the markers determines the kind of distance index which can be computed. Without genetic interpretation, genetic distance has to be computed on band information, and this has been shown to be less discriminant because of redundancy. Therefore, it is recommended that monolocus probes be chosen and that distance indices be computed using allelic information. With RFLP studies, data are generally available with different restriction enzymes for the same probe. The present study confirms that those different combinations of enzymes per probe are not independent. To avoid redundancy, it has generally been proposed to choose one enzyme per probe. As a matter of fact, information is optimized with one enzyme per probe. In this study however, three enzymes per probes have been tested. The most discriminant distance index was obtained by considering the combination of profiles obtained for different enzymes with the same probe as new alleles.

Relationship between molecular and morphological distances

Divergence on molecular markers behaves as a limiting factor of morphological divergence (Fig. 4). Two explanations can be proposed for this relationship:

1) First of all, it is clear from quantitative genetic theory that two different combinations of genes may lead to the same phenotype. This generates a triangular relationship between the distance for a quantitative trait and the proportion of quantitative trait loci (QTLs) involved in the variation of this trait and for which 2 inbred lines differ (Charcosset 1992). Similar relationships would be obtained for distances computed from several quantitative traits (which is the case of the Mahalanobis distance), provided each of them is controlled by several loci.
2) Secondly, since it is not possible to estimate directly the proportion of QTLs for which 2 inbred lines differ, distance is computed using genetic markers. Linkage disequilibrium between markers and QTLs affects the relationship between morphological distance and marker distance. If there is no linkage disequilibrium between markers and QTLs, the two distances will vary independently; high and low marker distance will correspond to similar morphological distance. Properties of (1) and (2) have been investigated by Burstin and Charcosset (1997). Both of these properties contribute to the general triangular tendency of Fig. 4.

Consequences for distinctness studies

In this paper we considered a set of ten morphological traits recorded at different stages of plant growth and combined them into a distance index. Experimental results concerning the relationship between marker and morphological distances illustrated the efficiency of this approach for germplasm protection: if 2 inbred lines differ at the morphological level, they will differ at the marker level. Thus, on this multiple-trait basis, the

probability that a line is registered, although very close (at the DNA level) to a pre-existing line, appears to be very low. However, it is clear that markers would bring complementary information in case of morphological similarity, allowing discrimination between: (1) a close similarity that is the result of different breeding sources (different combinations of genes), and (2) a close similarity due to copies or very high relatedness.

Acknowledgements We are grateful to all the persons who obtained the data used in this study, and especially to C. Faye, C. Johnsson and P. Tatout for RFLP data production; Z. Karaman for database management. We also thank all the participants of the "distinction working group" for helpful discussions, and one anonymous reviewer for his comments on the manuscript. This research was funded by the SEPROMA association and the French Ministry of Agriculture.

References

Bar-Hen A, Charcosset A (1995) Relationship between molecular and morphological distances in a maize inbred lines collection. Application for breeder's rights protection. In: Van Ooijen JW, Jansen J (eds) Biometrics in plant breeding: applications of molecular markers (Proc 9th Meet Eucarpia Sect Biometrics Plant Breed). CPRO-DLO, Wageningen, pp 57–66

Bar-Hen A, Charcosset A, Bourgoin M, Guiard J (1995) Relationship between genetic markers and morphological traits in a maize inbred lines collection. Euphytica 84:145–154

Bernardo R (1993) Estimation of coefficient of coancestry using molecular markers in maize. Theor Appl Genet 85:1055–1062

Bishop YMM, Fienberg SE, Holland PW (1975) Discrete multivariate analysis: theory and practice. The MIT Press, Cambridge

Burstin J, Charcosset A, Barrière Y, Hebert Y, De Vienne D, Damerval C (1995) Molecular markers and protein quantities as genetic descriptors in maize. II. Prediction of performance of hybrids for forage traits. Plant Breed 114:427–433

Burstin J, Charcosset A (1997) Relationship between phenotypic and marker distances: theoretical and experimental investigations. Heredity (in press)

Charcosset A (1992) Prediction of heterosis. In: Dattlée Y, Dumas C, Gallais A (eds) Reproductive biology and plant breeding. Springer, Berlin Heidelberg New York, pp 355–369

Cochran WG (1977) Sampling techniques, 3rd edn. Wiley, New York

Dubreuil P, Dufour P, Krejci E, Causse M, De Vienne D, Gallais A, Charcosset A (1996) Organization of RFLP diversity among inbred lines of Maize representing the most significant heterotic groups. Crop Sci 36:790–799

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

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

Gower JC (1985) Measures of similarity, dissimilarity and distance. In: Kotz S, Johnson NL (eds) Encyclopedia of statistical sciences, vol 5. Wiley, New York, pp 397–405

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

Mahalanobis PC (1936) On the generalized distances in statistics. Proc Nat Inst Sci India 2:49–55

Malecot G (1948) Les Mathématiques de l'hérédité. Masson et Cie, Paris

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

Melchinger AE, Messmer MM, Lee M, Woodman WL, Lamkey KR (1991) Diversity and relationships among U.S. Maize inbreds revealed by Restriction Fragment Length Polymorphisms. Crop Sci 31:669–678

Messmer MM, Melchinger AE, Lee M, Woodman WL, Lamkey KR (1991) Diversity and relationships among US maize inbreds revealed by restriction fragment length polymorphism. Crop Sci 31:669–678

Murigneux A, Barloy D, Leroy P, Beckert M (1993) Molecular and morphological evaluation of doubled haploid lines in maize. 1. Homogeneity within DH lines. Theor Appl Genet 86:837–842

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:3269–3273

Nei M, Roychoudhury AK (1973) Sampling variances of heterozygosity and genetic distance. Genetics 76:379–390

Rogers JS (1972) Measures of similarity and genetic distance. Studies in Genetics VII. Univ Texas Publ 7213:145–153

Smith JSC, Smith OS (1989) The description and assessment of distance between inbred lines of maize. I. The use of morphological traits as descriptors. Maydica 34:141–150

Smith JSC, Smith OS, Bowen SL, Tenborg RA, Wall SJ (1991a) The description and assessment of distances between lines of Maize. III. A revised scheme for the testing of distinctness between inbred lines utilizing DNA RFLPs. Maydica 36:213–226

Smith JSC, Smith OS, Wall SJ (1991b) Associations among widely used French and US maize hybrids as revealed by restriction fragment length polymorphisms. Euphytica 54:263–273

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

Soller M, Beckmann JS (1983) Genetic polymorphism in varietal identification and genetic improvement. Theor Appl Genet 47:179–190