Theor Appl Genet (2005) 111: 598–608
DOI 10.1007/s00122-005-2052-2

ORIGINAL PAPER

M. Heckenberger · M. Bohn · M. Frisch · H. P. Maurer
A. E. Melchinger

Identification of essentially derived varieties with molecular markers: an approach based on statistical test theory and computer simulations

Received: 23 December 2004 / Accepted: 21 April 2005 / Published online: 26 May 2005

Abstract Genetic similarities (GS) based on molecular markers have been proposed as a tool for identification of essentially derived varieties (EDVs). Nevertheless, scientifically reliable criteria for discrimination of EDVs and independently derived varieties with GS estimates are scanty, and implementation into practical breeding has not yet taken place. Our objectives were to (1) assess the influence of chromosome number and length, marker density, and distribution, as well as the degree of polymorphism between the parental inbreds on the distribution of GS between parental inbreds and their progenies [$GS_{(P1,O)}$] derived from F_2 and different backcross populations and (2) evaluate these factors with regard to the power for distinguishing F_2- versus BC_1- and BC_1- versus BC_2-derived lines with molecular markers. We developed an approach based on statistical test theory for the identification of EDVs with molecular markers. Standard deviations and overlaps of distributions of $GS_{(P1,O)}$ of F_2-, BC_1-, and BC_2-derived lines were smaller with (1) increasing chromosome number and length, (2) increasing marker density, and (3) uniformly instead of randomly distributed markers, approaching a lower boundary determined by the genetic parameters. The degree of polymorphism between the parental inbreds influenced the power only if the remaining number of polymorphic markers was low. Furthermore, suggestions are made for (1) determining the number of markers required to ascertain a given power and (2) EDV identification procedures.

Keywords Essentially derived varieties · Genetic similarity · Intellectual propert · Molecular markers · Parental contribution · Computer simulation

Communicated by R. Bernardo

M. Heckenberger and M. Bohn contributed equally to this manuscript.

M. Bohn (✉)
Crop Science Department, University of Illinois,
S-110 Turner Hall, 1102 South Goodwin Avenue,
Urbana, IL 61801, USA
E-mail: mbohn@uiuc.edu

M. Heckenberger · M. Frisch · H. P. Maurer · A. E. Melchinger
Institute of Plant Breeding, Seed Science,
and Population Genetics, University of Hohenheim,
70593 Stuttgart, Germany

Introduction

The convention act of the International Union for the Protection of New Varieties of Plants (UPOV) allows the use of protected germplasm for generating new, improved varieties, without authorization by the breeder of the protected variety (UPOV 1978). The goal of the so-called ''breeder's exemption'' is to secure future breeding progress and to prevent genetic erosion in elite breeding germplasm. Thus, plant breeders can use protected varieties as a source of initial variation to create new base populations and select for improved varieties in subsequent breeding steps. These new varieties earn protection if they comply with the UPOV criteria of distinctness, uniformity, and stability.

Application of genetic engineering techniques and molecular markers in breeding programs has provided potential opportunities to misuse the breeder's exemption in its original intention by adding only one or a few genes to a protected variety. To cope with this new situation and provide a basis for discussion of legal implications, the concept of essential derivation was added to the revised UPOV Convention Act (UPOV 1991). Accordingly, a variety is essentially derived from an initial variety, when ''(1) it is predominantly derived from the initial variety, or from a variety that is itself predominantly derived from the initial variety, while retaining the expression of the essential characteristics that result from the genotype or combination of genotypes of the initial variety; (2) it is clearly distinguishable from the initial variety; and (3) except for the differences which result from the act of derivation, it conforms to the initial variety in the expression of the essential

characteristics that result from the genotype or combination of genotypes of the initial variety.''

Whereas UPOV guidelines are in place to determine the distinctness among varieties, no regulations have been fixed for measuring conformity among varieties. In particular, the revised UPOV convention does not specify methods for determining the genetic conformity of an initial variety and putative essentially derived varieties (EDVs), but mentions only examples of breeding procedures that may lead to EDVs (e.g., ''backcrossing'' or selection of ''natural or induced mutants''). Consequently, it is currently up to the breeders to agree on breeding methods that yield EDVs and on methods for determining genetic conformity between an initial variety and potential EDVs. Therefore, breeding organizations such as the American Seed Trade Association (ASTA), the Association of French Maize Breeders, and the International Organization of Plant Breeders (ASSINSEL) are currently developing guidelines for the identification of EDVs in maize and other crops.

A document of the ASSINSEL (1999) demanded ''scientifically reliable criteria'' for identification of EDVs and, in addition to other methods, proposed the use of molecular markers to evaluate the degree of genetic conformity between initial variety and putative EDVs. Because molecular markers, such as simple sequence repeats (SSRs) or amplified fragment length polymorphisms, allow tracing chromosomal segments from parents to their progeny, genetic similarities (GS) based on molecular markers were regarded as suitable tools to distinguish EDVs from independently derived varieties (ASSINSEL 1999; International Seed Federation 2002). In addition to some pioneering studies describing both theoretical (Dillmann et al. 1997; Wang and Bernardo 2000) and empirical results (Bernardo and Kahler 2001), a detailed investigation of the factors influencing the distribution of GS between parental inbreds and their progenies derived with different breeding procedures is lacking. In addition, thresholds already suggested for the identification of lines developed by accepted or unaccepted breeding procedures must be evaluated critically before they are employed on a routine basis.

With the possibility of generating ''virtual'' breeding populations by computer simulations (Frisch et al. 2000), it has become feasible to determine the probability distribution of the contribution of a parental line to the genome of a derived line, depending on (1) the mating scheme used to generate the derived line and (2) the number and length of the chromosomes of the crop under consideration. Such distributions can be used to investigate the statistical power of marker-based tests, depending on marker position, density, and degree of polymorphism of markers in the parents.

The main goal of this study was to develop an approach for the identification of EDVs based on statistical test theory and to provide benchmark data for decisions on EDV thresholds in crops with various genome size. In detail, our objectives were to (1) assess the influence of chromosome number and lengths, genome coverage and distribution of markers, and the degree of polymorphism between the parental inbreds on the distribution of estimates of GS between parental inbreds and their progenies derived from F_2 and different backcross populations and (2) evaluate these factors with regard to the statistical power of molecular markers to discriminate F_2- versus BC_1- and BC_1- versus BC_2-derived lines.

Materials and methods

Assumptions

Suppose progeny line O was developed by single seed descent (SSD) without selection from an F_2, BC_1, or BC_2 population obtained from a cross of homozygous parental lines P1 and P2. For backcross-derived lines, P1 is the recurrent parent. GS between lines P1, P2, and O are denoted by $GS_{(P1,P2)}$, $GS_{(P1,O)}$, and $GS_{(P2,O)}$. They are obtained as GS = 1−GD, where GD is the Rogers distance (Rogers 1972) or any other distance measure fulfilling the criteria given by Melchinger (1993). Accuracy of estimated GS values was determined by using the root mean square error,

$$\sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\mathrm{GS}} - \mathrm{GS}_{\mathrm{true}}\right)^2},$$

where $\widehat{\mathrm{GS}}$ denotes the GS estimated by molecular markers, GS_{true} denotes the true GS, which remains unknown in practice but is known in our computer simulations (see below), and n denotes the sample size.

If GS is determined from a large number of marker loci with uniform coverage of the genome, we have (Heckenberger et al. 2005)

$$\mathrm{GS}_{(\mathrm{P1,O})} = \mathrm{GS}_{(\mathrm{P1,P2})} + p\left(1 - \mathrm{GS}_{(\mathrm{P1,P2})}\right), \quad (1)$$

where p is the parental genome contribution from P1 to O. Solving Eq. 1 for p yields the estimate

$$\hat{p} = \frac{\mathrm{GS}_{(\mathrm{P1,O})} - \mathrm{GS}_{(\mathrm{P1,P2})}}{1 - \mathrm{GS}_{(\mathrm{P1,P2})}}. \quad (2)$$

Further formulas for the calculation of $\mu_{GS_{(P1,O)}}$ and the variance $\sigma^2_{GS_{(P1,O)}}$ were given by Heckenberger et al. (2005). As only markers polymorphic between P1 and P2 can be used to estimate $\hat{p}$, we define the ''effective number of markers'' (m_e) as

$$m_e = m * \left(1 - \mathrm{GS}_{(\mathrm{P1,P2})}\right), \quad (3)$$

where m is the total number of markers applied.

Likewise, the ''effective marker distance'' (d_e) can be calculated in centiMorgans as

$$d_e = \bar{d} * \frac{1}{1 - \mathrm{GS}_{(\mathrm{P1,P2})}}, \quad (4)$$

where $\bar{d}$ is the average map distance between adjacent markers of the marker set applied. Furthermore, m_e can be defined as the total genome length divided by d_e, which results in the formula

$$m_e = \frac{l_t}{d_e}, \quad (5)$$

where l_t is the total length of the genome of the particular crop in centiMorgans.

Knowledge of the distribution of $GS_{(P1,O)}$ for defined pedigree relationships between P1 and O is a key prerequisite to develop a statistical test for identifying EDVs. Hitherto, analytical expressions of this distribution are unknown. In the absence of selection, the distribution of $GS_{(P1,O)}$ depends on the pedigree and the number and length of the chromosomes. If P1 and P2 are unrelated [coefficient of coancestry $f_{(P1,P2)} = 0$, Malécot 1948], then the expectation $\mu_{GS_{(P1,O)}}$ is equal to $f_{(P1,O)}$ and, thus, $\mu_{GS_{(P1,O)}} = 0.500$, 0.750, and 0.875 for F_2-, BC_1-, or BC_2-derived lines, respectively. In addition, formulas to calculate the variance of p for F_2- and BC_1-derived lines were given by Wang and Bernardo (2000).

Simulation studies

With computer simulations using the software Plabsoft (Maurer et al. 2004), we determined approximated distributions of $GS_{(P1,O)}$ for F_2-, BC_1-, and BC_2-derived SSD lines, assuming that P1 and P2 are completely homozygous. The simulation of each crossing scheme was repeated 50,000 times to reduce sampling effects and to obtain high numerical accuracy of the results.

Simulations were performed for various scenarios differing in the following parameters:

1. Number of chromosomes: five (*Arabidopsis thaliana* L.), ten (maize, *Zea mays* L.), and 20 chromosomes [similar to sunflower (*Helianthus anuus*) 17 chromosomes, oilseed rape (*Brassica napus* L.) 19 chromosomes, and bread wheat (*Triticum aestivum* L.) 21 chromosomes].
2. Chromosome length: 20, 40, 80, 160, and 320 cM.
3. Average marker distance: 4, 8, 16, 32, 64, 128, or 256 cM.
4. GS between the parental inbreds $GS_{(P1,P2)} = 0.00$, 0.25, 0.50, or 0.75, with a random location of monomorphic markers in the parents.
5. Distribution of markers: random versus uniform.

The locations of uniformly distributed markers were chosen according to the nonterminal model of Wang and Bernardo (2000), where the distance between either end of the chromosome and the first marker is half the distance between the markers. Only scenarios with at least two markers per chromosome were analyzed. To study the effects of chromosome number and length, all chromosomes within a given scenario were of equal length, which is in some crop species only a crude approximation of reality.

Statistical test

We consider inbred lines P1, P2, and O, for which the marker genotypes are known, and assume that development of inbred lines from an F_2 population by SSD is an accepted breeding procedure. Inbred lines having a significantly greater similarity to P_1 than expected for F_2-derived lines are assumed to be derived from a backcross generation, the latter being considered as EDVs. Note that identical principles apply to other scenarios if, for example, derivation of lines from a BC_1 population is considered an accepted breeding procedure but the use of higher backcross generations not. We test the null hypothesis.

- H_0: Line O is an F_2-derived inbred line against the alternative hypothesis.
- H_A: Line O is more closely related to P1 than expected for F_2-derived inbred lines.

$\widehat{GS}_{(P1,O)}$ is used as test statistic and H_0 is accepted if $\widehat{GS}_{(P1,O)}$ is smaller than the $1-\alpha$ percentile T of the distribution of $\widehat{GS}_{(P1,O)}$ under H_0 obtained from simulations. If $\widehat{GS}_{(P1,O)} > T$, H_0 is rejected.

For this test, the type I error α is the probability that line O is an F_2-derived inbred line, but the test incorrectly suggests the rejection of H_0. The type II error β is the probability that line O is more closely related to P1 than an F_2-derived inbred line [$f_{(P1,O)} > 0.5$], but the test incorrectly suggests the acceptance of H_0. The power $1-\beta$ of the test to detect an EDV depends on the value chosen for α and the generations to which the test is applied (e.g., F_2- versus BC_1-derived or BC_1- versus BC_2-derived).

In our investigations, we started with an ideal situation and stepwise approached reality by removing simplifying assumptions. First, we assumed an infinite number of markers with uniform genome coverage and 100% polymorphism between the parental inbred lines [$GS_{(P1,P2)} = 0$]. Second, a reduction in the marker density and a shift from uniform to random marker distribution in the genome were investigated under the assumption of $\widehat{GS}_{(P1,P2)} = 0$. Third, the parental inbreds were not completely polymorphic, but $\widehat{GS}_{(P1,P2)} > 0$ with a random distribution of monomorphic markers among markers uniformly covering the entire genome.

Results

Genetic similarities across the whole genome

For all simulated scenarios of plant genomes that were characterized by different chromosome numbers and lengths, the mean parental contributions to the genome of the progeny were consistent with the expected p values for each breeding generation, i.e., 0.500 for F_2-, 0.750 for

BC_1-, and 0.875 for BC_2-derived progenies. Here, p reflects the "true" genetic similarity between P1 and O under the assumption of an infinite number of markers and fully polymorphic parental lines. Standard deviations (SDs) for true p values decreased from F_2- to BC_2- derived lines and from smaller to larger genomes (Table 1). For the same total genome length, SDs were smaller for genomes with shorter, but more, chromosomes than for those with longer, but fewer, chromosomes. Therefore, 90, 95, and 99% percentiles were smaller for larger genomes. In addition, the distribution of p showed a negative kurtosis for F_2-, and a positive kurtosis for BC_2-derived lines. The kurtosis for BC_1-derived lines was close to zero for all scenarios examined. The kurtosis for larger genomes with longer or more chromosomes was generally closer to zero than for smaller genomes. For the same total genome length, kurtosis was closer to zero for genomes with shorter, but more, chromosomes than for genomes with longer, but fewer, chromosomes.

The highest theoretically achievable power $1-\beta$ of the test based on $\widehat{GS}_{(P1,O)}$ to distinguish between F_2- versus BC_1- and BC_1- versus BC_2-derived lines was evaluated by examining the extent of overlaps in the distributions of p for F_2-, BC_1-derived lines (Fig. 1), as well as for BC_1- versus BC_2-derived lines (data not shown). Given a type I error (α) of 0.05 for F_2-derived lines, the power 1-β increased with the total genome size (Table 2). For the same total genome length, 1-β was higher for genomes with more, but shorter, chromosomes compared to genomes with fewer, but longer, chromosomes. These trends also held true for a discrimination of BC_1- versus BC_2-derived lines. However, 1-β decreased drastically for advanced backcross generations.

Marker estimates of genetic similarities

The root mean square error ($\sqrt{MSE}$) for marker-based estimates of genetic similarity ($\widehat{GS}_{(P1,O)}$) increased with a reduction in the number of markers per chromosome (Table 3). For the same marker density, $\sqrt{MSE}$ decreased from F_2 to BC_2 generations, from longer to shorter chromosomes, and from 5 to 20 chromosomes.

For a constant number of markers per chromosome, correlations [$r_{(p;GS)}$] between the true parental contribution p and $\widehat{GS}_{(P1,O)}$ dropped considerably with increasing genome length, especially for lower marker densities (Table 4). In contrast, $r_{(p;GS)}$ was fairly stable across different chromosome lengths for a constant distance between the markers. In addition, $r_{(p;GS)}$ slightly decreased from F_2- to BC_2- derived lines, but remained almost constant across different numbers of chromosomes.

Critical thresholds T for $\widehat{GS}_{(P1,O)}$ on the basis of $\alpha = 0.05$ increased with decreasing marker density (Table 2). Compared with the theoretically achievable power, 1-β decreased only marginally for all evaluated genomes with dense marker maps and marker distances below 16 cM. With a further reduction in marker density, 1-β dropped substantially. For example, for a scenario similar to the maize genome (ten chromosomes of 160-cM length) and $\alpha = 0.05$, the power 1-β for F_2- versus BC_1-derived lines amounted to 0.91 and decreased only marginally for marker distances up to 16 cM, whereas it dropped to 0.64 for BC_1- versus BC_2-derived lines. This loss of power 1-β for the test in advanced backcross generations compared with F_2- versus BC_1-derived lines was higher for smaller than for larger genomes, especially with short chromosomes.

Marker coverage and polymorphism

If polymorphic markers are randomly distributed across the genome, $r_{(p;GS)}$ was generally lower compared with a uniform distribution of markers across the genome (Table 3). This reduction in $r_{(p;GS)}$ was higher with

Table 1 Standard deviation (*SD*) and kurtosis, as well as 90, 95, and 99% percentiles of the distributions of the true parental genome contribution (p) to offspring lines derived by singe-seed descent from segregating F_2, BC_1, and BC_2 populations of biparental crosses of homozygous parents, for various genome parameters, ordered by total genome length

Genome parameters			SD			Kurtosis			Percentile								
Total genome length (cM)	Chromosomes		F_2	BC_1	BC_2	F_2	BC_1	BC_2	F_2			BC_1			BC_2		
	No.	Length							90%	95%	99%	90%	95%	99%	90%	95%	99%
400	5	80	0.157	0.130	0.096	−0.267	−0.001	0.667	0.704	0.760	0.856	0.912	0.946	0.990	0.990	1.000	1.000
400	10	40	0.128	0.108	0.081	−0.159	−0.036	0.345	0.666	0.712	0.795	0.885	0.916	0.967	0.974	0.992	1.000
400	20	20	0.100	0.085	0.064	−0.093	−0.035	0.173	0.628	0.664	0.730	0.857	0.883	0.929	0.952	0.971	0.997
800	5	160	0.127	0.103	0.076	−0.193	0.001	0.549	0.665	0.710	0.790	0.879	0.908	0.952	0.966	0.981	1.000
800	10	80	0.111	0.092	0.068	−0.148	−0.009	0.335	0.643	0.683	0.752	0.865	0.892	0.936	0.958	0.973	0.995
800	20	40	0.091	0.076	0.057	−0.069	−0.026	0.186	0.617	0.650	0.710	0.846	0.870	0.911	0.945	0.961	0.984
1,600	5	320	0.097	0.078	0.058	−0.115	−0.004	0.370	0.625	0.660	0.723	0.848	0.871	0.911	0.945	0.959	0.981
1,600	10	160	0.090	0.073	0.054	−0.099	0.002	0.251	0.616	0.648	0.706	0.842	0.865	0.902	0.941	0.955	0.976
1,600	20	80	0.078	0.065	0.048	−0.073	0.005	0.161	0.601	0.629	0.681	0.832	0.852	0.888	0.934	0.948	0.969
3,200	10	320	0.069	0.055	0.041	−0.073	0.002	0.175	0.588	0.613	0.658	0.820	0.838	0.869	0.925	0.937	0.956
3,200	20	160	0.064	0.052	0.038	−0.048	0.001	0.114	0.582	0.605	0.648	0.815	0.832	0.862	0.922	0.934	0.952

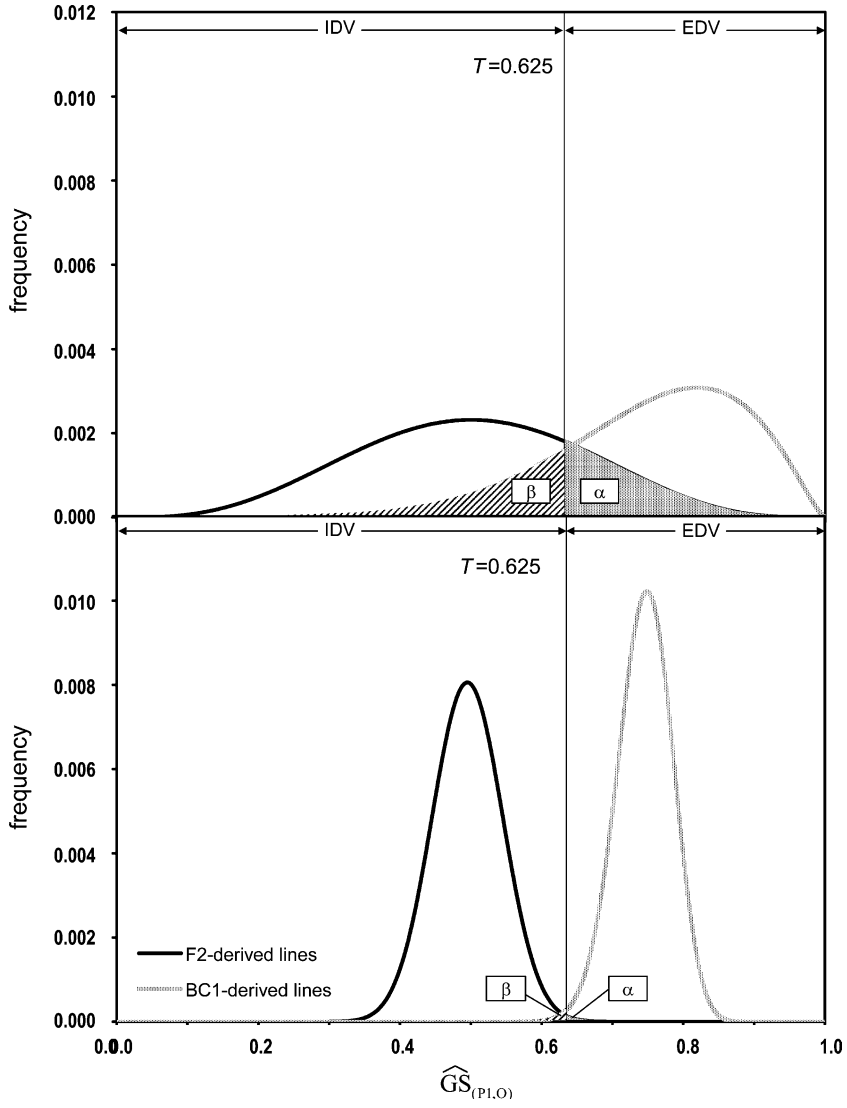

Fig. 1 Illustration of type I (α) and type II (β) errors made by distinguishing F_2- and BC_1-derived lines by genetic similarity ($\widehat{GS}_{(P1,O)}$) between F_2-, or BC_1- derived lines and their (recurrent) parental line under the assumption of a critical threshold (T) of $\widehat{GS}_{(P1,O)} = 0.625$ to distinguish between essentially (*EDV*) and independently (*IDV*) derived varieties. Frequency distributions of $\widehat{GS}_{(P1,O)}$ were obtained by 50,000 simulation runs, assuming a uniform marker distance of 4 cM and either five chromosomes of 80-cM length (*upper diagram*), or 20 chromosomes of 320-cM length (*lower diagram*)

decreasing marker density and chromosome length, and more pronounced for genomes with fewer chromosomes. For a constant total genome length and a marker number proportional to the genome length, the reduction was smaller for genomes with fewer, but longer, chromosomes. For a given average marker distance, 1-β was generally smaller for randomly than for uniformly distributed markers (data not shown).

Assuming incomplete polymorphism among the parental inbreds for the applied set of markers ($\widehat{GS}_{(P1,P2)} > 0$), estimates of $r_{(p;GS)}$ dropped considerably with increasing $GS_{(P1,P2)}$, especially for lower marker densities. However, if the marker density was sufficiently high ($d_e < 16$ cM), $r_{(p;GS)}$ decreased only marginally with increasing values of $GS_{(P1,P2)}$. Increasing $\widehat{GS}_{(P1,P2)}$ from 0 to higher values had only little effect on the power 1-β of the test to distinguish between F_2- versus BC_1- and BC_1- versus BC_2-derived lines, as long as the marker density of polymorphic markers (m_e) remained high (Fig. 2). In addition, the power 1-β remained fairly constant for a given effective marker distance (d_e), if $GS_{(P1,P2)}$ was ≤ 0.5, but decreased for $\widehat{GS}_{(P1,P2)} > 0.5$.

Discussion

Means and variances of simulated frequency distributions of the parental contribution to F_2- and BC_1-derived lines fitted closely to the values obtained by the formulas of Wang and Bernardo (2000) for the nonterminal marker model. In addition, the ratio of σ^2_p for BC_1- and F_2-derived lines was exactly 0.75, as predicted by theory.

In the literature, accuracy of GS values is commonly determined by calculating standard errors (SEs) with bootstrap or jackknife methods (Efron 1982), using markers as sample unit (cf. Dreisigacker et al. 2004; Lombard et al. 2000; Reif et al. 2004). However, these resampling methods require stochastically independent

Table 2 Thresholds (T) and discriminatory power ($1-\beta$) for a given type I error (α) of 0.05 between F_2- vs BC_1- and BC_1- vs BC_2-derived lines based on true parental genome contributions (p) and marker-based genetic similarities ($\widehat{GS}_{(P1,O)}$) depending on marker distance, chromosome number, and chromosome length, assuming uniformly distributed markers and $GS_{(P1,P2)}=0$

Total genome length	Chromosomes		T							$1-\beta$						
	No.	Length (cM)	Marker distance (cM)							Marker distance (cM)						
			0[a]	4	8	16	32	64[b]	128[b]	0[a]	4	8	16	32	64[b]	128[b]
F_2 vs BC_1																
400	5	80	0.75	0.76	0.77	0.77	0.78	0.81	0.86	0.52	0.53	0.50	0.50	0.46	0.43	0.36
400	10	40	0.71	0.72	0.73	0.72	0.74	0.75	–	0.67	0.64	0.61	0.62	0.58	0.54	–
400	20	20	0.66	0.67	0.67	0.67	0.69	–	–	0.84	0.82	0.81	0.80	0.77	–	–
800	5	160	0.70	0.71	0.72	0.72	0.74	0.75	0.79	0.68	0.66	0.66	0.64	0.58	0.54	0.47
800	10	80	0.68	0.69	0.68	0.69	0.70	0.72	0.76	0.78	0.75	0.75	0.75	0.69	0.65	0.53
800	20	40	0.65	0.65	0.65	0.66	0.66	0.69	–	0.90	0.88	0.89	0.85	0.83	0.75	–
1,600	5	320	0.66	0.65	0.66	0.66	0.67	0.69	0.73	0.87	0.88	0.85	0.85	0.82	0.73	0.61
1,600	10	160	0.64	0.65	0.65	0.66	0.66	0.68	0.70	0.91	0.90	0.89	0.88	0.84	0.77	0.70
1,600	20	80	0.62	0.63	0.63	0.64	0.64	0.65	0.68	0.96	0.95	0.95	0.93	0.91	0.87	0.76
3,200	10	320	0.61	0.61	0.61	0.61	0.62	0.64	0.66	0.99	0.99	0.99	0.98	0.98	0.93	0.84
3,200	20	160	0.60	0.60	0.61	0.61	0.62	0.63	0.64	1.00	0.99	0.99	0.99	0.98	0.95	0.92
BC_1 vs BC_2																
400	5	80	0.93	0.94	0.93	0.94	0.95	0.96	0.98	0.34	0.32	0.35	0.33	0.31	0.27	0.26
400	10	40	0.90	0.91	0.91	0.91	0.92	0.94	–	0.41	0.41	0.42	0.39	0.37	0.36	–
400	20	20	0.87	0.88	0.88	0.88	0.90	–	–	0.55	0.52	0.51	0.50	0.47	–	–
800	5	160	0.90	0.90	0.91	0.91	0.92	0.93	0.95	0.42	0.41	0.42	0.40	0.39	0.35	0.32
800	10	80	0.88	0.89	0.89	0.90	0.90	0.91	0.94	0.50	0.47	0.47	0.46	0.42	0.41	0.33
800	20	40	0.86	0.87	0.87	0.87	0.88	0.89	–	0.61	0.60	0.59	0.57	0.55	0.47	–
1,600	5	320	0.86	0.87	0.87	0.87	0.88	0.90	0.92	0.60	0.60	0.58	0.58	0.53	0.44	0.35
1,600	10	160	0.86	0.86	0.86	0.87	0.88	0.89	0.90	0.64	0.62	0.63	0.60	0.55	0.50	0.44
1,600	20	80	0.85	0.85	0.85	0.86	0.86	0.87	0.90	0.72	0.71	0.69	0.67	0.63	0.59	0.46
3,200	10	320	0.83	0.84	0.83	0.84	0.84	0.86	0.87	0.84	0.81	0.83	0.81	0.77	0.67	0.54
3,200	20	160	0.83	0.83	0.83	0.83	0.84	0.85	0.86	0.88	0.86	0.85	0.83	0.79	0.71	0.64

[a]The true parental contribution was used to determine the theoretically highest achievable discriminatory power
[b]At least two markers per chromosome were used for the simulations

sample units as an assumption, which is violated for linked markers, especially when the marker density is high. To circumvent this problem, we calculated $\sqrt{\text{MSE}}$ as a measure for the accuracy of estimated GS values, because this requires no assumptions about sampling units. Calculation of $\sqrt{\text{MSE}}$ is only possible in computer simulations, when the true GS values are known. In practice, however, correct calculation of SEs for estimated GS values represents an unsolved problem and warrants further research.

For a map with terminal markers, Wang and Bernardo (2000) found that the SE of parental contribution at marker loci reached a minimum with approximately 200 markers, even for large genomes. This is in contrast with our study, where $\sqrt{\text{MSE}}$ approached zero with an increasing number of markers, which was in agreement with theoretical expectations (Dubreuil et al. 1996; Tivang et al. 1994). We conjecture that this discrepancy in the results between the two studies is partly attributable to different assumptions concerning interference. Our simulations were based on the absence of interference, as assumed for Haldane's (1919) mapping function, whereas Wang and Bernardo (2000) applied the assumptions of Kosambi's (1944) mapping function, which adjusts for interference of recombination events.

The increasing skewness and kurtosis from F_2- to BC_2-derived lines can be explained by the fact that the distribution of p is restricted between the boundaries of 0 and 1. This leads to more skewed distributions, the closer the mean approaches the boundaries. In addition, the 90, 95, and 99% percentiles of the distribution of p observed for maize genome parameters (ten chromosomes of 160 cM) were slightly lower than experimental data published by Bernardo and Kahler (2001), who examined the parental contribution of a large set of F_2-derived maize inbreds with 60 SSRs and 20 restriction fragment length polymorphisms. This can be explained by the relatively small number of markers used by these authors.

Effect of chromosome number and length on the power $1-\beta$

We observed that genomes with larger and more chromosomes showed smaller values of σ^2_p and, therefore, a higher power $1-\beta$ of the statistical test (Table 2). As recombination events occur more often in larger than in smaller genomes, this results in a larger number of chromosomal segments transferred from parents to the progeny in large genomes, particularly with a large number of chromosomes. Therefore, σ^2_p decreases with increasing genome length as a consequence of the central limit theorem, which states that the distribution of a sum

Table 3 Average root mean square errors ($\overline{\sqrt{\text{MSE}}}$) of marker-estimated genetic similarities ($\widehat{\text{GS}}_{(\text{P1,O})}$) between F_2-, BC_1-, and BC_2-derived lines and their (recurrent) parent, depending on chromosome length and marker distance

Chromosome length (cM)	Markers per chromo-some	Marker distance (cM)	$\overline{\sqrt{MSE}}$								
			F_2			BC_1			BC_2		
			5 Chromosomes	10 Chromosomes	20 Chromosomes	5 Chromosomes	10 Chromosomes	20 Chromosomes	5 Chromosomes	10 Chromosomes	20 Chromosomes
80	2	40	0.083	0.058	0.041	0.076	0.054	0.038	0.060	0.042	0.030
	3	27	0.059	0.042	0.029	0.054	0.038	0.027	0.043	0.030	0.021
	5	16	0.038	0.027	0.019	0.035	0.025	0.017	0.028	0.019	0.014
	10	8	0.020	0.014	0.010	0.018	0.013	0.009	0.014	0.010	0.007
	20	4	0.010	0.007	0.005	0.009	0.007	0.005	0.007	0.005	0.004
160	2	80	0.105	0.074	0.052	0.095	0.067	0.047	0.075	0.053	0.037
	3	54	0.075	0.053	0.038	0.069	0.049	0.034	0.054	0.038	0.027
	5	32	0.048	0.034	0.024	0.045	0.032	0.022	0.035	0.025	0.018
	10	16	0.026	0.018	0.013	0.024	0.017	0.012	0.019	0.013	0.009
	20	8	0.013	0.009	0.007	0.012	0.009	0.006	0.010	0.007	0.005
	40	4	0.007	0.005	0.003	0.006	0.004	0.003	0.005	0.003	0.002
320	2	160	0.124	0.088	0.062	0.111	0.079	0.056	0.086	0.061	0.043
	3	107	0.093	0.065	0.046	0.084	0.059	0.042	0.065	0.046	0.033
	5	64	0.062	0.043	0.031	0.056	0.040	0.028	0.044	0.031	0.022
	10	32	0.033	0.024	0.017	0.031	0.022	0.015	0.024	0.017	0.012
	20	16	0.017	0.012	0.009	0.016	0.011	0.008	0.013	0.009	0.006
	40	8	0.009	0.006	0.004	0.008	0.006	0.004	0.007	0.005	0.003
	80	4	0.004	0.003	0.002	0.004	0.003	0.002	0.003	0.002	0.002

Table 4 Correlations between true parental genome contributions and GS, depending on chromosome number, chromosome length, and average marker distance. A uniform (*u*) or random (*r*) distribution of markers and $GS_{(P1,P2)} = 0$ between the parental inbreds of the biparental cross was assumed

Chromosome length	Markers per chromosome	Marker distance (cM)	Marker distribution	F_2			BC_1			BC_2		
				5 Chromosomes	10 Chromosomes	20 Chromosomes	5 Chromosomes	10 Chromosomes	20 Chromosomes	5 Chromosomes	10 Chromosomes	20 Chromosomes
80	2	40	u	0.900	0.900	0.900	0.881	0.884	0.880	0.869	0.866	0.867
			r	0.712	0.721	0.709	0.602	0.618	0.606	0.517	0.566	0.610
80	3	27	u	0.944	0.944	0.944	0.932	0.933	0.932	0.923	0.922	0.923
			r	0.780	0.774	0.773	0.685	0.689	0.652	0.618	0.646	0.683
80	5	16	u	0.975	0.975	0.976	0.970	0.969	0.969	0.966	0.966	0.966
			r	0.855	0.852	0.845	0.764	0.749	0.688	0.717	0.748	0.743
80	10	8	u	0.993	0.993	0.993	0.991	0.991	0.991	0.990	0.990	0.990
			r	0.918	0.912	0.915	0.833	0.808	0.746	0.825	0.823	0.801
80	20	4	u	0.998	0.998	0.998	0.998	0.998	0.998	0.997	0.997	0.997
			r	0.958	0.954	0.954	0.870	0.837	0.766	0.880	0.869	0.830
320	2	160	u	0.652	0.652	0.651	0.609	0.610	0.612	0.588	0.581	0.589
			r	0.519	0.517	0.518	0.419	0.446	0.414	0.371	0.410	0.427
320	3	107	u	0.749	0.751	0.755	0.712	0.709	0.710	0.689	0.687	0.689
			r	0.610	0.604	0.597	0.530	0.505	0.473	0.443	0.486	0.494
320	5	64	u	0.857	0.857	0.860	0.826	0.829	0.829	0.811	0.811	0.809
			r	0.695	0.701	0.701	0.612	0.574	0.517	0.573	0.594	0.566
320	10	32	u	0.950	0.949	0.949	0.935	0.936	0.934	0.926	0.926	0.926
			r	0.817	0.811	0.806	0.700	0.638	0.559	0.719	0.695	0.633
320	20	16	u	0.985	0.985	0.985	0.980	0.980	0.980	0.977	0.977	0.977
			r	0.891	0.882	0.890	0.765	0.704	0.592	0.799	0.750	0.702
320	40	8	u	0.996	0.996	0.996	0.995	0.995	0.995	0.994	0.994	0.994
			r	0.942	0.935	0.938	0.800	0.715	0.604	0.847	0.799	0.715
320	80	4	u	0.999	0.999	0.999	0.999	0.999	0.999	0.998	0.998	0.998
			r	0.967	0.967	0.969	0.815	0.739	0.631	0.874	0.825	0.749

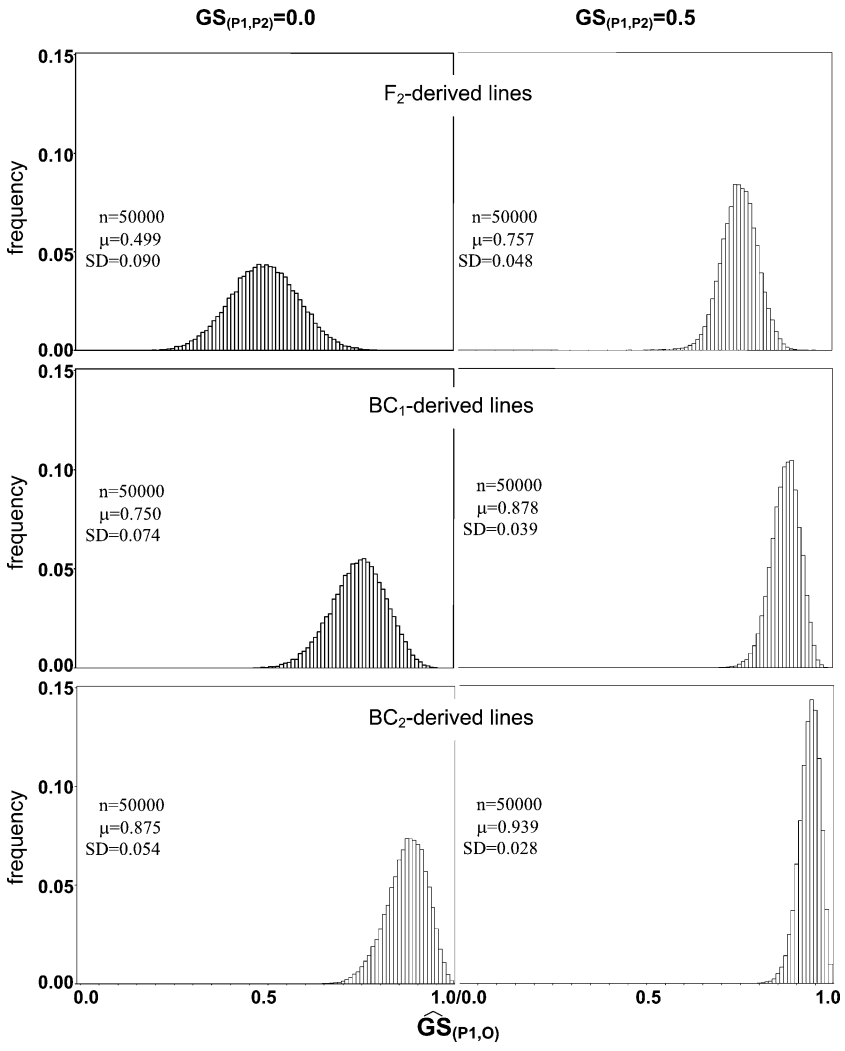

Fig. 2 Histograms of genetic similarities ($\widehat{GS}_{(P1,O)}$) between F_2-, BC_1-, or BC_2-derived lines and their recurrent parent, assuming a genetic similarity between the parental lines [$GS_{(P1,P2)}$] of 0.0 or 0.5. The latter was generated by randomly choosing markers monomorphic between the parents. Simulations were based on 50,000 runs, assuming ten chromosomes of 160-cM length and 40 uniformly distributed markers per chromosome

of independent random variables (i.e., p as a function of the sum of chromosomal segments transferred from P to O) approaches a normal distribution with increasing sample size (i.e., a larger number of chromosomal segments in genomes with large chromosome numbers).

The number of chromosomes was found to have the largest impact on σ^2_p and, therefore, on the power $1-\beta$ of the statistical test based on $\widehat{GS}_{(P1,O)}$ to distinguish between F_2- versus BC_1-derived inbreds, was found for chromosome number. Doubling of the chromosome number led to a maximum gain of power $1-\beta$ of up to 50% for a constant chromosome length, whereas doubling of the chromosome length led to a considerably smaller gain in power. Likewise, if scenarios with the same total genome length were compared, genomes with shorter and more chromosomes showed a higher power $1-\beta$ than genomes with longer, but fewer, chromosomes. The reason for this is that an increase in the chromosome number leads to more independent linkage groups with independent recombination events. Especially for very short chromosomes, the effect of more linkage groups on the number of independently inherited chromosomal segments is higher than an increase in the genome length for a given number of linkage groups, because the latter only affects recombination events within already existing linkage groups. This implies that the best theoretically achievable power $1-\beta$ of $\widehat{GS}_{(P1,O)}$ to distinguish lines derived from F_2, BC_1, or BC_2 populations is generally higher for crops with larger genomes and many chromosomes (e.g., sunflower, wheat, oilseed rape) than for crops with smaller genomes and few chromosomes, e.g., barley (*Hordeum vulgare* L.) or rye (*Secale cereale* L.). Consequently, these large effects of chromosome number and length on the discriminatory power require crop-specific thresholds for EDV identification.

Marker density and distribution

Correlations [$r_{(p;GS)}$] between true parental contributions and marker-estimated values for $GS_{(P1,O)}$ were fairly constant for a given marker distance, irrespective of the number and length of chromosomes, but decreased with increasing chromosome length for a constant number of

markers per chromosome. Obviously, what matters for $r_{(p;GS)}$ is that each marker represents the same proportion of the genome, i.e., a constant marker distance.

Using randomly instead of uniformly distributed markers has the same effect in that it reduces the marker coverage of the genome. Consequently, a careful selection of the marker set applied for the identification of EDVs is highly recommended (Dreher et al. 2003; Morris et al. 2003). This will save expenditures in the lab assays, because in the case of a random marker distribution, at least twice as many markers are necessary to reach the same power $1-\beta$ for $\widehat{GS}_{(P1,O)}$ as for uniformly distributed markers. If the initial set of markers is uniformly distributed, but their degree of polymorphism low, the remaining polymorphic markers among the parental inbreds will no longer be uniformly distributed across the genome. Both factors (nonuniform marker distribution and low degree of polymorphism of the markers) lead to nonrepresented parts of the genome. Thus, the number of markers must be increased accordingly to cover the remaining parts of the genome with polymorphic markers. Solving Eq. 3 for m and inserting the mean $\widehat{GS}_{(P1,P2)}$ of unrelated lines of the particular crop as well as the desirable m_e determined with Eq. 5 on the basis of a given d_e may serve as a rule of thumb for estimating the necessary number of markers to reach a given power $1-\beta$. In addition, a correction factor accounting for a nonuniform marker distribution would be desirable, which warrants further research. It should be used in Eqs. 3 and 4 in a way that the number of markers m actually applied yields the same $r_{(p;\widehat{GS}_{(P1,O)})}$ as with m_e uniformly distributed markers and $\widehat{GS}_{(P1,P2)} = 0$.

Especially for crops with low degrees of polymorphism (e.g., barley), the choice of highly polymorphic markers is very important to avoid a loss of power due to high numbers of noninformative (monomorphic) markers. Therefore, the set of markers should be optimized by selecting a set of highly polymorphic markers with the restrictions of a constant number of markers per unit of chromosome length, e.g., at least two markers per chromosomal bin. This should be feasible for all major crops as in most cases a high number of SSRs is already available or currently being developed.

If the effective marker distance d_e is 16 cM or lower, a further increase in the marker density leads only to a minor increase in the power $1-\beta$, which approaches the maximum power achievable when applying the true parental contribution p determined with an infinite number of markers. This is in close agreement with results published by Wang and Bernardo (2000). Nevertheless, marker-based GS values have a certain estimation error that is a function of the number of markers and of considerable size, even when d_e is 20 cM or lower (Heckenberger et al. 2005). As a consequence, a confidence interval for the true GS of one or two SEs around $\widehat{GS}_{(P1,O)}$ estimated by markers should be applied. This would reduce the type I error, but also considerably decrease the power $1-\beta$ of the test. Given (1) an upper limit for $\sqrt{MSE}$ or SE of $\widehat{GS}_{(P1,O)}$, (2) the mean degree of polymorphism of the applied marker system in the crop, and (3) the power $1-\beta$ to be achieved, one can calculate the necessary number of markers by formulas of Eeuwijk and Baril (2001) and Foulley and Hill (1999). In addition, the application of genetic similarity measures that consider information of marker positions, as suggested by Dillmann et al. (1997) might also help to account for nonuniform marker distributions and lead to more precise GS estimates.

Guidelines for EDV identification procedures

If both parental inbreds of a putative EDV are known, we recommend the following procedure. First, $\widehat{GS}_{(P1,P2)}$ and $\widehat{GS}_{(P1,O)}$ must be determined by using a sufficient number of molecular markers. Second, $\hat{p}$ (i.e., $\widehat{GS}_{(P1,O)}$ on the basis of the markers polymorphic between the parents) can be estimated with the aid of Eq. 2 under omittance of nonparental bands. Alternatively, solutions for including nonparental bands (Bernardo et al. 2000) can be applied. Third, a critical threshold T can be determined based on a given α or $1-\beta$ using simulated distributions of $\widehat{GS}_{(P1,O)}$ on the basis of the polymorphic markers of the same genetic map that was applied for the marker assay, i.e., generating a large number of virtual F_2-, and BC_1-derived lines from the same cross analyzed with the same markers. Finally, the test decision can be made by comparing $\widehat{GS}_{(P1,O)}$ of the putative EDV with T. If $\widehat{GS}_{(P1,O)}$ is beyond the critical threshold T, a reversal of the burden of proof would occur (Eeuwijk and Law 2004), and the breeder of the putative EDV would have to supply evidence (e.g., breeding books) that accepted breeding procedures were used. This approach has the advantage that it is independent of the initial degree of relatedness of the parental inbreds. In addition, it can be adapted to arbitrary crop- and marker-system-specific parameters. Therefore, our results can be easily adapted to genomes of other diploid or allopolyploid crops not explicitly mentioned in this study.

In a proposal to ASTA, an estimated parental genome contribution of $p \geq 0.75$ suggests an EDV (Smith et al. 1991). The reasoning for this threshold reflects the viewpoint of some commercial breeding companies that regard backcrossing as an unacceptable breeding procedure in maize. Bernardo and Kahler (2001) studied the range of parental contribution among maize lines developed from F_2 populations without and with selection. They concluded that inbreds with 70% to nearly 80% of their genome derived from one parent could be obtained from an F_2 population. Our results show that a threshold of $p = 0.75$ applied for a genome similar to maize would result in very low values of α (0.001 for ≥ 200 polymorphic markers). However, it is associated

with a power $1-\beta$ of only 0.50, because decreasing values of α are necessarily associated with increasing values of β. In comparison, a fixed α of 0.05 (instead of a fixed T) would result in $T=0.64$ and $1-\beta$ of 0.91. The choice of an appropriate T then depends more or less on the kind of politics that is pursued. Implementation of thresholds with a very high power $1-\beta>0.90$ would avoid backcrossing, but result in high values of α (i.e., large numbers of independently derived varieties falsely judged as EDVs), whereas thresholds with low α would result in a large number of lines that were bred by unaccepted breeding procedures but not judged as EDV.

If only one parent is known, it is not possible to determine $\widehat{GS}_{(P1,P2)}$ directly. Therefore, the test decision can only be based on the distribution of $\widehat{GS}_{(P1,O)}$. As this distribution depends on the unknown $\widehat{GS}_{(P1,P2)}$, one possibility is to use the mean $(\mu_{\widehat{(GS_{(P1,P2)})}})$ between unrelated lines of the respective germplasm instead of $\widehat{GS}_{(P1,P2)}$. Obviously, this adjustment would be too conservative if the parents P1 and P2 have $\widehat{GS}_{(P1,P2)} > \mu_{\widehat{(GS_{(P1,P2)})}}$, i.e., they are more similar than expected for unrelated lines, but too liberal for very distant parents. However, these deviations would be in the spirit of the revised UPOV convention that intends to avoid plagiarism and encourages efforts to enhance the genetic diversity of elite breeding germplasm.

Acknowledgements We are indebted to the Gesellschaft zur Förderung der privaten deutschen Pflanzenzüchtung e.V. (GFP), Germany, for a grant supporting M. Bohn. Financial support for M. Heckenberger was provided by the European Union, grant no. QLK-CT-1999-01499 (MMEDV).

References

ASSINSEL (1999) Essential derivation and dependence. Practical information. http://www.worldseed.org/Position_papers/derive.htm

Bernardo R, Kahler AL (2001) North American study on essential derivation in maize: inbreds developed without and with selection from F_2 populations. Theor Appl Genet 102:986–992

Bernardo R, Romero-Severson J, Ziegle J, Hauser J, Joe L, Hookstra G, Doerge RW (2000) Parental contribution and coefficient of coancestry among maize inbreds: pedigree, RFLP, and SSR data. Theor Appl Genet 100:552–556

Dillmann C, Charcosset A, Goffinet B, Smith JSC, Dattée Y (1997) Best linear unbiased estimator of the molecular genetic distance between inbred lines. Advances in biometrical genetics. In: Krajewski P, Kaczmarek Z (eds) Proceedings of the tenth meeting of the EUCARPIA section biometrics in plant breeding, 14–16 May 1997, Poznan, Poland, pp 105–110

Dreher K, Khairallah M, Ribaut JM, Morris M (2003) Money matters (I): costs of field and laboratory procedures associated with conventional and marker-assisted maize breeding at CIMMYT. Mol Breed 11:221–234

Dreisigacker S, Zhang P, Warburton ML, Van Ginkel M, Hoisington D, Bohn M, Melchinger AE (2004) SSR and pedigree analyses of genetic diversity among CIMMYT wheat lines targeted to different megaenvironments. Crop Sci 44:381–388

Dubreuil P, Dufour P, Krejci E, Causse M, Devienne 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

Eeuwijk FAv, Baril CP (2001) Conceptual and statistical issues related to the use of molecular markers for distinctness and essential derivation. Acta Hortic 546:35–53

Eeuwijk FAv, Law JR (2004) Statistical aspects of essential derivation, with illustrations based on lettuce and barley. Euphytica 137:129–137

Efron B (1982) The bootstrap, the jackknife, and other resampling plans. SIAM, Philadelphia

Foulley JL, Hill WG (1999) On the precision of estimation of genetic distance. Genet Sel Evol 31:457–464

Frisch M, Bohn M, Melchinger AE (2000) Plabsim: software for simulation of marker-assisted backcrossing. J Hered 91:86–87

Haldane JBS (1919) The combination of linkage values, and the calculation of distances between loci of linked factors. J Genet 8:299–309

Heckenberger M, Bohn M, Melchinger AE (2005) Identification of essentially derived varieties obtained from biparental crosses of homozygous lines. I. SSR data from maize inbreds. Crop Sci 45:1132–1140

International Seed Federation (2002) ISF view on intellectual property. International Seed Federation, Chicago

Kosambi DD (1944) The estimation of map distances from recombination values. Ann Eugen 12:172–175

Lombard V, Baril CP, Dubreuil P, Blouet F, Zhang D (2000) Genetic relationships and fingerprinting of rapeseed cultivars by AFLP: consequences for varietal registration. Crop Sci 40:1417–1425

Malécot G (1948) Les mathematiques de l'heredité. Masson & Cies, Paris

Maurer HP, Melchinger AE, Frisch M (2004) Plabsoft: software for simulation and data analysis in plant breeding. Proceedings of the 17th Eucarpia General Congress, 8–11 September 2004, Tulln, Austria, pp 359–362

Melchinger AE (1993) Use of RFLP markers for analysis of genetic relationships among breeding materials and prediction of hybrid performance. In: Buxton DR, et al (ed) International Crop Science I. CSSA, Madison, pp 621–628

Morris M, Dreher K, Ribaut JM, Khairallah M (2003) Money matters (I): costs of maize inbred line conversion schemes at CIMMYT using conventional and marker-assisted selection. Mol Breed 11:235–247

Reif JC, Xia XC, Melchinger AE, Warburton ML, Hoisington DA, Beck D, Bohn M, Frisch M (2004) Genetic diversity determined within and among CIMMYT maize populations of tropical, subtropical, and temperate germplasm by SSR markers. Crop Sci 44:326–334

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

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

Tivang JG, Nienhuis J, Smith OS (1994) Estimation of sampling variance of molecular marker data using the bootstrap procedure. Theor Appl Genet 89:259–264

UPOV (1978) International convention for the protection of new varieties of plants. http://www.upov.int/en/publications/conventions/1978/content.htm

UPOV (1991) International convention for the protection of new varieties of plants. http://www.upov.int/en/publications/conventions/1991/content.htm

Wang JK, Bernardo R (2000) Variance of marker estimates of parental contribution to F_2 and BC_1-derived inbreds. Crop Sci 40:659–665