Aus dem Institut für
Pflanzenzüchtung, Saatgutforschung und Populationsgenetik
der Universität Hohenheim
Fachgebiet Angewandte Genetik und Pflanzenzüchtung
Prof. Dr. A.E. Melchinger

Assessing the Genetic Diversity in Crops with Molecular Markers: Theory and Experimental Results with CIMMYT Wheat and Maize Elite Germplasm and Genetic Resources

Dissertation
zur Erlangung des Grades eines Doktors
der Agrarwissenschaften vorgelegt
der Fakultät Agrarwissenschaften

von
Diplom-Agrarbiologe
Jochen Reif
aus Gernsbach

Stuttgart-Hohenheim
2004

Die vorliegende Arbeit wurde am 04. Mai 2004 von der Fakultät Agrarwissenschaften als "Dissertation zur Erlangung des Grades eines Doktors der Agrarwissenschaften (Dr. sc. agr.)" angenommen.

Tag der mündlichen Prüfung: 15. September 2004
1. Prodekan: Prof. Dr. K. Stahr
Berichterstatter, 1. Prüfer: Prof. Dr. A.E. Melchinger
Mitberichterstatter, 2. Prüfer: Prof. Dr. H.-P. Piepho
3. Prüfer: Prof. Dr. R. Blaich

Contents

---

[1] Reif, J.C., A.E. Melchinger, and M. Frisch. 2004. Crop Science. In press.

[2] Reif, J.C., P. Zhang, S. Dreisigacker, M.L. Warburton, M. van Ginkel, D. Hoisington, M. Bohn, and A.E. Melchinger. 2004. Proc. Natl. Acad. Sci. USA. In review.

[3] Reif, J.C., X.C. Xia, A.E. Melchinger, M.L. Warburton, D.A. Hoisington, D. Beck, M. Bohn, and M. Frisch. 2004. Crop Science 44:326–334.

[4] Reif, J.C., A.E. Melchinger, X.C. Xia, M.L. Warburton, D.A. Hoisington, S.K. Vasal, G. Srinivasan, M. Bohn, and M. Frisch. 2003. Crop Science 43:1275–1282.

[5] Reif, J.C., A.E. Melchinger, X.C. Xia, M.L. Warburton, D.A. Hoisington, S.K. Vasal, D. Beck, M. Bohn, and M. Frisch. 2003. Theor. Appl. Genet. 107:947–957.

Abbreviations

AFLP	amplified fragment length polymorphism
ANOVA	analysis of variance
AMOVA	analysis of molecular variance
CIMMYT	International Maize and Wheat Improvement Center
COP	coefficient of parentage
GCA	general combining ability
HWE	Hardy-Weinberg equilibrium
LC	landrace cultivars
LD	linkage disequilibrium
ME	mega-environment
MRD	Modified Rogers distance
MWC	modern wheat cultivars
OTU	operational taxonomic unit
PCoA	principal coordinate analysis
PMPH	panmictic midparent heterosis
Pop	population
QTL	quantitative trait locus
RAPD	random amplified length polymorphism
RD	Rogers' distance
RFLP	restriction fragment length polymorphism
SCA	specific combining ability
SNP	single nucleotide polymorphism
SSR	simple sequence repeat

1. General Introduction

As the human population is steadily growing and the arable land is decreasing, the world faces a greater demand on agricultural output than ever before in history (Lee, 1998). In the past, this demand for increased agricultural productivity was met by a combination of genetic improvements, cultivation of more land, increased water supply, enhanced fertilization, use of pesticides, advanced mechanization, and favorable socioeconomic conditions (Tanksley and McCouch, 1997).

But as (i) freshwater reserves and petroleum resources, on which fertilizers and pesticides are based, are dwindling and (ii) problems caused by agricultural pollution are increasing, the current levels of agricultural inputs can hardly be enhanced or even maintained. Similarly, the existing farmland is decreasing due to urban and industrial development or natural phenomena such as expanding deserts. This leaves the genetic improvement of crops as the most viable and sustainable approach by which food production can attempt to keep pace with the anticipated growth of the human population (Hoisington et al., 1999).

For the genetic approach to succeed, the genetic variation provided by nature and currently conserved in seed banks must be harnessed. The seed bank collections as a source of genetic diversity must be well-characterized for efficient management and effective exploitation. The advent of PCR-based molecular markers, such as simple sequence repeats (SSRs), has created an opportunity for fine-scale genetic characterization of germplasm collections.

Molecular markers can be used for (i) detection of relationships among different germplasm in seed banks, (ii) search for promising heterotic groups for hybrid breeding, (iii) identification of duplicates in seed banks, and (iv) assessment of the level of genetic diversity present in germplasm pools and its flux over time.

In these various applications, a proper choice of a similarity s or dissimilarity coefficient $d = 1 - s$ (following the terminology of Gower, 1985) is important and depends on factors such as (i) the properties of the marker system employed, (ii) the genealogy of the germplasm, (iii) the operational taxonomic unit under consideration (*e.g.*, lines, populations), (iv) the objectives of the study, and (v) the necessary preconditions for subsequent multivariate analysis.

In a recent review, Mohammadi and Prasanna (2003) discussed the use of six coefficients d for the analysis of dichotomous molecular marker data, but ignored those coefficients based on allele frequencies, which are especially suitable for codominant marker data. Several authors (Goodman, 1972; Gower, 1985; Gower and Legendre, 1986) investigated the mathematical properties and relationships among various coefficients d. Nevertheless, coefficients were disregarded, which are based on specific genetic models. However, in particular these coefficients are suitable for studies with seed bank or plant breeding materials.

For an efficient characterization of germplasm with molecular markers with special focus on applications in plant breeding and seed banks, a thorough review of the genetical and mathematical properties of coefficients d is required. Such a review has not yet been compiled and published.

Flux of Diversity in Wheat

Wheat belongs to the genus *Triticum* that originated in the historic Fertile Crescent, an area in the Middle East, almost 10 000 years ago. *Triticum*

arose from the cross of two diploid wild grasses, resulting in tetraploid wheat (*T. turgidum* L.) (Salamini et al., 2002). Tetraploid wheats later crossed to diploid goat grasses (*T. tauschii*) and gave rise to hexaploid wheat (*T. aestivum* L.), also known as bread wheat (Kihara, 1944; McFadden and Sears, 1946). This hexaploid wheat has been considered the product of just a few independent crosses between its progenitors (Dvorak et al., 1998; Talbert et al., 1998). A loss of diversity from the two original forms, *T. dicoccoides* and *T. tauschii* resulting in hexaploid wheat, is presumably due to the limited number of crosses involved in this evolutionary process.

Through the centuries, mutation generated new alleles, while recombination created novel allele combinations. This genetic variation was subsequently reduced by (i) genetic drift and (ii) natural and early farmer selection, resulting in a series of landraces adapted to the specific conditions of their habitats.

During the last century, traditional landraces of most crop plants have been continually replaced by modern or high yielding crop varieties. These modern varieties were bred with a limited number of landraces in their pedigree and it is postulated that they contained less genetic diversity than landraces (Frankel, 1970). Thus, a popular hypothesis is that modern plant breeding and intensive selection over an extended period have further reduced genetic diversity among cultivars (Tanksley and McCouch, 1997). Such reduction may have consequences both on the vulnerability of crops to pests and on their ability to respond to changes in climate or agricultural practice (FAO, 1998). The first signs that germplasm with a narrow genetic base might lead to disasters in wheat came from several severe epidemics of shoot fly (*Atherigona* spp.) and karnal bunt (*Tilletia indica*) in India in the 1970s (Dalrymple, 1986).

During the last 40 years, the International Maize and Wheat Improvement Center (CIMMYT) has had a large impact on spring wheat. In all developing countries excluding China, approximately 86% of the spring bread wheat area in 1997 was sown with CIMMYT or CIMMYT-related germplasm involving at least one CIMMYT ancestor (Smale et al., 2002). Therefore, CIMMYT's

wheat germplasm is exceptionally suitable for investigation whether domestication and breeding have reduced genetic diversity in wheat in detrimental manner. This information can help in broadening the genetic base of the elite breeding pool by introgression of landraces and/or wild ancestors of wheat. Nevertheless, an in-depth study of the diversity trends during the domestication and breeding of wheat is still lacking.

Population Structure of Maize Germplasm

From 1964 until 1973 CIMMYT's breeding program developed and improved a wide array of maize populations each of which was derived from one single racial complex. In 1974, germplasm from different racial complexes was mixed and more than 100 populations were established to use the combining ability of different germplasm sources for intra-population improvement. In addition, 30 broad-based back-up pools were formed as an insurance against narrowing the germplasm base of the populations (CIMMYT, 1998). The intermixing of diverse germplasm within populations complicates a detection of relationships among these populations based on pedigree information.

One possible approach to detect genes and alleles of interest in germplasm collections is association mapping (Lynch and Walsh, 1997). This could be a strategy for a systematic exploitation of the diversity present in CIMMYT's germplasm. The resolution of association studies in a sample depends on the extent of linkage disequilibrium (LD) across the genome. LD (or the correlation between alleles of different loci) generally depends on the genealogy of the germplasm. Moreover, drift and selection within populations can also cause LD. The genomic structure of LD must be empirically determined before conducting association studies, because it varies among samples of germplasm.

Detailed knowledge about LD and genetic diversity of the CIMMYT populations is required to increase the efficiency of their use in breeding. Never-

theless, information about (i) molecular diversity in tropical and subtropical maize populations and (ii) LD in this germplasm is entirely lacking.

Establishment of Heterotic Groups

Assigning germplasm to heterotic groups and patterns is fundamental in hybrid breeding for a maximum exploitation of heterosis (Hallauer et al. 1988). While heterotic patterns in temperate maize were established more than 50 years ago, a clearly defined heterotic pattern does not exist in tropical and subtropical maize of the CIMMYT germplasm. Therefore, before initiating a hybrid breeding program, CIMMYT conducted several diallel studies to determine heterotic relationships among CIMMYT populations and pools. Several of the populations demonstrated good general combining ability, and various promising heterotic patterns were identified (Beck et al., 1990; Beck et al., 1991; Crossa et al., 1990; Vasal et al., 1992a,b,c). However, no conclusions were drawn about clearly defined heterotic groups, because of the mixed genetic constitution of the populations.

Lamkey and Edwards (1999) coined the term panmictic midparent heterosis (PMPH) to describe the deviation in performance between a population cross and the mean of its two parent populations in Hardy-Weinberg equilibrium. Quantitative genetic theory shows that in the absence of epistasis and two alleles per locus, PMPH is a function of the product of the dominance effect and the square of the difference in gene frequencies at the respective locus (Falconer and Mackay, 1996, p. 255), which corresponds to the square of the modified Rogers distance (Melchinger, 1999). In fact, a linear increase in PMPH with increasing genetic distance (Hypothesis 1) was observed in a diallel of U.S. maize populations (Moll et al., 1962).

In contrast, a study with tropical maize populations (Moll et al., 1965) of diverse geographic origin suggest that PMPH increases with increasing

genetic distance only up to an optimum level but thereafter decreases in extremely wide crosses (Hypothesis 2). The authors explained their results by fertility distortion in wide crosses and negative epistatic interactions between unadapted genes. While Moll et al. (1962, 1965) inferred the genetic distance from the geographic origin of the populations, to our knowledge no attempts have been made to verify or discard the above hypotheses with more reliable distance measures based on molecular markers.

If heterosis of hybrids increases monotonically with increasing genetic distance of the parents, then genetic distances based on molecular markers are a useful tool for establishing promising heterotic groups and patterns. Nevertheless, up to date no thorough analysis of the relationship between genetic distances and heterosis is available.

Objectives

The goal of my thesis research was to optimize the conservation and use of CIMMYT's genetic resources with the aid of molecular markers. In particular, the objectives were to

1. investigate the genetical and mathematical properties of 10 dissimilarity coefficients widely used in germplasm surveys and determine relationships between these coefficients;

2. examine consequences of the properties of the coefficients on different areas of application in plant breeding and seed banks;

3. examine the loss of genetic diversity during (i) domestication of bread wheat, (ii) transition from traditional landraces to modern wheat breeding varieties, and (iii) intensive selection over a sustained period of international wheat breeding;

4. investigate the molecular genetic diversity within and among 23 CIMMYT maize populations with the use of SSR markers;

5. examine genotype frequencies for deviations from Hardy-Weinberg equilibrium at individual loci and test for linkage disequilibrium between pairs of loci;

6. investigate the relationship of SSR-based genetic distances between populations and panmictic midparent heterosis in a broad range of CIMMYT maize germplasm;

7. evaluate the usefulness of SSR markers for defining heterotic groups and patterns in tropical and subtropical maize germplasm; and

8. examine applications of SSR markers for broadening heterotic groups by systematic introgression of other germplasm.

References

Beck, D.L., S.K. Vasal, and J. Crossa. 1990. Heterosis and combining ability of CIMMYT's tropical early and intermediate maturity maize (*Zea mays* L.) germplasm. Maydica 35:279–285.

Beck, D.L., S.K. Vasal, and J. Crossa. 1991. Heterosis and combining ability among subtropical and temperate intermediate-maturity maize germplasm. Crop Sci. 31:68–73.

CIMMYT. 1998. A complete listing of maize germplasm from CIMMYT. Maize Program Special Report. Mexico DF, Mexico.

Crossa, J., S.K. Vasal, and D.L. Beck. 1990. Combining ability estimates of CIMMYT tropical late yellow maize germplasm. Maydica 35:273–278.

Dalrymple, D.G. 1986. Development and spread of high-yield wheat varieties in developing countries. Bureau for Science and Technology, U.S. Agency for International Development, Washington, DC.

Dvorak, J., M.C. Luo, Z.L. Yang, and H.B. Zhang. 1998. The structure of the *Aegilops tauschii* genepool and the evolution of hexaploid wheat. Theor. Appl. Genet. 67:657–670.

Falconer, D.S., and T.F.C. Mackay. 1996. Introduction to Quantitative Genetics. 4th ed. Longman Group Ltd., London.

Food and Agricultural Organization (FAO). 1998. The state of the world's plant genetic resources for food and agriculture. FAO, Rome.

Frankel, O.H. 1970. Genetic dangers of the Green Revolution. World Agric. 19:9–14.

Goodman, M.M. 1972. Distance analysis in biology. Syst. Zool. 174–186.

Gower, J.C. 1985. Measures of similarity, dissimilarity and distances. p. 397–405. *In* S. Kotz, N.L. Johnson, and C.B. Read (eds.) Encyclopedia of Statistical Sciences, Vol. 5. Wiley, Ney York.

Gower, J.C., and P. Legendre. 1986. Metric and Euclidean properties of dissimilarity coefficients. J. Classification 3:5–48.

Hallauer, A.R., W.A. Russell, and K.R. Lamkey. 1988. Corn breeding. p. 463–564. *In* G.F. Sprague and J.W. Dudley (eds.) Corn and Corn Improvement. 3rd ed. Agron. Monogr. 18. ASA, CSSA, and SSSA, Madison, WI.

Hoisington, D., M. Khairallah, T. Reeves, J-M. Ribaut, B. Skovmand, S. Taba, and M.L. Warburton. 1999. Plant genetic resources: What can they contribute toward increased crop productivity. Proc. Natl. Acad. Sci. USA 96:5937–5943.

Kihara, H. 1944. Die Entdeckung der DD-Analysatoren beim Weizen. Agr. and Hort. (Tokyo) 19:889-890.

Lamkey, K.R., and J.W. Edwards. 1999. Quantitative genetics of heterosis. p. 31–48. *In* J.G. Coors and S. Pandey (eds.) The Genetics and Exploitation of Heterosis in Crops. CSSA, Madison, WI.

Lee, M. 1998. Genome projects and gene pools: new germplasm for plant breeding. Proc. Natl. Acad. Sci. USA 90:5095–5099.

Lynch, M., and B. Walsh. 1997. Genetics and Analysis of Quantitative Traits. p. 413. Sinauer Assoc., Sunderland, MA.

McFadden, E.S., and E.R. Sears. 1946. The origin of *Triticum spelta* and its free threshing hexaploid relatives. J. Hered. 37:81-89.

Melchinger, A.E. 1999. Genetic diversity and heterosis. p. 99–118. *In* J.G. Coors and S. Pandey (eds.) The Genetics and Exploitation of Heterosis in Crops. CSSA, Madison, WI.

Mohammadi, S.A., and B.M. Prasanna. 2003. Analysis of genetic diversity in crop plants – salient statistical tools and considerations. Crop Sci. 43:1235–1248.

Moll, R.H., W.S. Salhuana, and H.F. Robinson. 1962. Heterosis and genetic diversity in variety crosses of maize. Crop Sci. 2:197–198.

Moll, R.H., J.H. Longquist, J.V. Fortuna, and E.C. Johnson. 1965. The relation of heterosis and genetic divergence in maize. Genetics 52:139–144.

Salamini, F., H. Özkan, A. Brandolini, R. Schäfer-Pregl, and W. Martin. 2002. Genetics and geography of wild cereal domestication in the near east. Nat. Rev. Genet. 3:429–441.

Smale, M., M.P. Reynolds, M.L. Warburton, B. Skovmand, R. Trethowan, R.P. Singh, I. Ortiz-Monasterio, and J. Crossa. 2002. Dimension of diversity in modern spring bread wheat in developing countries from 1965. Crop Sci. 42:1766–1779.

Tanksley, S.D., and S.R. McCouch. 1997. Seed banks and molecular maps: unlocking genetic potential from the wild. Science 277:1063–1066.

Talbert, L.E., L.Y. Smith, and N.K. Blake. 1998. More than one origin of hexaploid wheat is indicated by sequence comparison of low-copy DNA. Genome 41:402–407.

Vasal, S.K., G. Srinivasan, D.L. Beck, J. Crossa, S. Pandey, and C. de Leon. 1992a. Heterosis and combining ability of CIMMYT's tropical late white maize germplasm. Maydica 37:217–223.

Vasal, S.K., G. Srinivasan, J. Crossa, and D.L. Beck. 1992b. Heterosis and combining ability of CIMMYT's subtropical and temperate early maturity maize germplasm. Crop Sci. 32:884–890.

Vasal, S.K., G. Srinivasan, S. Pandey, H.S. Cordova, G.C. Han, and F. Gonzalez. 1992c. Heterotic patterns of ninety-two white tropical CIMMYT maize lines. Maydica 37:259–270.

Genetical and Mathematical Properties of Similarity and Dissimilarity Coefficients Applied in Plant Breeding and Seed Bank Management

J.C. Reif, A.E. Melchinger, and M. Frisch

Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany

Abstract

A proper choice of a dissimilarity measure is important in surveys investigating genetic relationships among germplasm with molecular marker data. The objective of our study was to examine 10 dissimilarity coefficients widely used in germplasm surveys, with special focus on applications in plant breeding and seed banks. In particular, we (i) investigated the genetical and mathematical properties of these coefficients, (ii) examined consequences of these properties for different areas of application in plant breeding and seed banks, and (iii) determined relationships between these 10 coefficients. The genetical and mathematical concepts of the coefficients were described in detail. A Procrustes analysis of a published data set consisting of seven CIMMYT maize populations demonstrated close affinity between Euclidean, Rogers', modified Rogers', and Cavalli-Sforza and Edwards' dissimilarity on one hand and Nei's standard and Reynolds dissimilarity on the other hand. Our investigations show that genetical and mathematical properties of dissimilarity measures are of crucial importance when choosing a genetic dissimilarity coefficient for analyzing molecular marker data. The presented results assist experimenters to extract the maximum amount of information from genetic data and, thus, facilitate the interpretation of findings from molecular marker studies on a theoretically sound basis.

QUANTIFYING the degree of dissimilarity among genera, species, subspecies, populations, and elite breeding materials is of primary concern in population genetics and plant breeding. Before 1970, measures of genetic dissimilarity between taxonomic units relied on pedigree analysis and morphological, physiological or cytological markers, as well as biometric analyses of quantitative and qualitative traits, heterosis or the segregation variance in crosses (Melchinger, 1999). During the following two decades, isozymes have successfully been employed in numerous taxonomic and evolutionary studies (Hamrick and Godt, 1997) but their use in other applications was hampered by the small number of polymorphic markers available.

Molecular markers, such as restriction fragment length polymorphisms (RFLPs), random amplified polymorphic DNA (RAPDs), amplified fragment length polymorphisms (AFLPs), simple sequence repeats (SSRs), and single nucleotide polymorphisms (SNPs) have meanwhile replaced isozymes and are heavily used for (i) detection of relationships among different germplasm in seed banks and breeding programs (c.f., Brummer, 1999), (ii) prediction of heterosis (c.f., Melchinger, 1999), (iii) search for promising heterotic groups for hybrid breeding (c.f., Reif et al., 2003), (iv) identification of duplicates in seed banks (c.f., Treuren et al., 2001), (v) assessment of the level of genetic diversity present in germplasm pools and its flux over time (c.f., Dubreuil and Charcosset, 1998; Labate et al., 2003), and (vi) identification of essentially derived varieties in plant variety protection (c.f., Smith et al., 1991; Lombard et al., 2000).

In these various applications, a proper choice of a similarity s or dissimilarity coefficient $d = 1 - s$ (following the terminology of Gower, 1985) is important and depends on factors such as (i) the properties of the marker system employed, (ii) the genealogy of the germplasm, (iii) the operational taxonomic unit (OTU) under consideration (e.g., lines, populations), (iv) the objectives of the study, and (v) necessary preconditions for subsequent multivariate analyses.

In a recent review, Mohammadi and Prasanna (2003) discussed the use of six coefficients d for the analysis of dichotomous molecular marker data, but ignored coefficients based on allele frequencies, which are especially suitable for codominant marker data. Several authors (Goodman, 1972; Gower, 1985; Gower and Legendre, 1986) investigated the mathematical properties and relationships among various coefficients d. However, those coefficients were disregarded, which are based on specific genetic models and, therefore, suitable for studies with seed bank or plant breeding materials.

In order to successfully conduct molecular marker surveys with plant breeding and seed bank materials, a thorough knowledge of genetical and mathematical properties of coefficients d is of crucial importance. Therefore, the objective of our study was to examine 10 coefficients d widely used in germplasm surveys, with special focus on applications in plant breeding and seed banks. In particular, we (i) investigated the genetical and mathematical properties of these coefficients, (ii) examined consequences of these properties for different areas of application in plant breeding and seed banks, and (iii) determined relationships between these 10 coefficients.

Corresponding author: Albrecht E. Melchinger, Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany. Email: melchinger@uni-hohenheim.de. Received January 28, 2004.

Table 1 Dissimilarity coefficients d for allelic informative marker data. p_{ij} and q_{ij} are allele frequencies of the jth allele at the ith locus in the two operational taxonomic units under consideration, n_i is the number of alleles at the ith locus, and m refers to the number of loci.

Variable	Dissimilarity coefficient		Range	Property	
				Distance	Euclidean
d_E	$\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(p_{ij}-q_{ij})^2}$	Euclidean	$0,\sqrt{2m}$	Yes	Yes
d_R	$\frac{1}{m}\sum_{i=1}^{m}\sqrt{\frac{1}{2}\sum_{j=1}^{n_i}(p_{ij}-q_{ij})^2}$	Rogers (1972)	0, 1	Yes	No
d_W	$\frac{1}{\sqrt{2m}}\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(p_{ij}-q_{ij})^2}$	Modified Rogers' (Wright, 1978; Goodman and Stuber, 1983)	0, 1	Yes	Yes
d_{CE}	$\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(1-\sum_{j=1}^{n_i}\sqrt{p_{ij}q_{ij}}\right)}$	Cavalli-Sforza and Edwards (1967)	0, 1	Yes	Yes
d_{RE}	$-\ln\left(1-\sum_{i=1}^{m}(a-b)/c\right)$ $a=1/2\sum_{j=1}^{n_i}(p_{ij}-q_{ij})^2$ $b=(1/(2(2n-1)))\times\left(2-\sum_{j=1}^{n_i}(p_{ij}^2+q_{ij}^2)\right)$ $c=\sum_{i=1}^{m}\left(1-\sum_{j=1}^{m}p_{ij}q_{ij}\right)$	Reynolds et al. (1983)	$0,\infty$	No	No
d_{N72}	$-\ln\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}p_{ij}q_{ij}}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n_i}p_{ij}^2\sum_{i=1}^{m}\sum_{j=1}^{n_i}q_{ij}^2}}$	Nei (1972)	$0, \infty$	No	No
d_{N83}	$\frac{1}{m}\sum_{i=1}^{m}\left(1-\sum_{j=1}^{n_i}\sqrt{p_{ij}q_{ij}}\right)$	Nei et al. (1983)	0, 1	No	No

Nature of molecular marker data

We suggest the term "allelic informative" if allele frequencies can be determined from the molecular marker data. Marker data are denoted as "allelic non-informative" if this is not feasible. For instance, SSR data of individual genotypes are allelic informative. AFLP data are mostly allelic non-informative although Geerlings et al. (1999), Piepho and Koch (2000), and Jansen et al. (2001) described methods to estimate allele frequencies and, thus, score AFLP data as allelic informative in specific situations.

Provided that molecular marker data are allelic informative, the estimates of coefficients d between OTUs can be calculated from the difference in the allele frequencies (Table 1) . For allelic non-informative molecular marker data, coefficients d based on absence or presence of observation of bands or signals must be applied (Table 2).

Distance and Euclidean Properties

Consider a set of elements M and a function d: $M \times M \rightarrow R$, assigning a real number to each pair of elements in M. A dissimilarity d is called a distance or metric, if for each element $i, j, k \in M$ the following three properties hold true (Gower, 1985):

$$d(i, j) \geq 0 \text{ and } d(i, j) = 0 \text{ if and only if } i = j, \quad (1)$$

$$d_{ij} = d_{ji}, \quad (2)$$

$$d_{ik} \leq d_{ij} + d_{jk}. \quad (3)$$

Some simple but important properties follow from this definition. All elements of a distance matrix with respect to a set of OTUs S must be defined and positive or zero. The matrix is symmetric and the triangle inequality (Equation 3) holds true for all triplets ($i, j, k \in$ S). The latter means that the length of any side of a triangle constructed with the three elements ($i, j, k \in$ S) is less than or equal to the sum of the lengths of the other two sides, with equality occurring only when the triangle degenerates to a line.

The coefficient d is Euclidean if n points $P_i \in R^n$ exist such that the Euclidean distance between P_i and P_j is d_{ij} for all $i, j \in M$ (Gower and Legendre, 1986). An illustration of the Euclidean property is given by these authors.

Table 2 Similarity coefficients for allelic non-informative marker data, where v_{ij} refers to the bands in common between two operational taxonomic units (OTUs) i and j, w_{ij} is the number of bands present in i and absent in j, x_{ij} is the number of bands present in j and absent in i, and y_{ij} is the number of bands both absent in i and j.

Variable	Similarity coefficient		Range	Property			
				Distance	Euclidean	Distance	Euclidean
				$1-s$		$\sqrt{1-s}$	
s_{SM}	$\frac{v_{ij}+y_{ij}}{v_{ij}+w_{ij}+x_{ij}+y_{ij}}$	Simple matching	0, 1	Yes	No	Yes	Yes
s_J	$\frac{v_{ij}}{v_{ij}+w_{ij}+x_{ij}}$	Jaccard (1908)	0, 1	Yes	No	Yes	Yes
s_D	$\frac{2v_{ij}}{2v_{ij}+w_{ij}+x_{ij}}$	Dice (1945)	0, 1	No	No	Yes	Yes

The Euclidean property is important, because it is an explicit or implicit assumption of many multivariate analysis methods such as principal coordinate analysis also known as classical multidimensional scaling, hierarchical cluster analysis, classification, hierachical classification, and graph theory (Gower, 1985). However, if a coefficient d is not Euclidean, then there exists a constant b greater than some minimal value such that the matrix with the elements $(d_{ij}+b)$ is Euclidean (Cox and Cox, 2000). The problem of finding such a constant b has been referred to for many years, Messick and Abelson (1956) being an early reference. Thus, the Euclidean property is desirable but the main criteria for the choice of a coefficient d are its genetical properties. Both, the Euclidean and genetical properties will be investigated for the coefficients d (Tables 1 and 2).

Genetic Dissimilarity Coefficients for allelic informative marker data

Euclidean Distance. The Euclidean distance is defined as:

$$d_E = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(p_{ij}-q_{ij})^2}, \quad (4)$$

where p_{ij} and q_{ij} are allele frequencies of the jth allele at the ith locus in the two OTUs under consideration, n_i is the number of alleles at the ith locus, and m refers to the number of loci. d_E ranges from zero to $\sqrt{2m}$, the limits being assumed when the two OTUs have identical allele frequencies or are fixed for different alleles. Thus, an obvious disadvantage is that d_E values from different studies cannot be compared directly because d_E depends on the number of marker loci assayed.

d_E is appropriate if allelic informative marker data are available and the relationships between OTUs (populations or individuals) are investigated in combination with multivariate methods that require dissimilarities possessing the Euclidean property.

Rogers' Distance. Rogers' distance (1972) is a modification of d_E and was developed assuming no knowledge about evolutionary forces diverging the OTUs under consideration:

$$d_R = \frac{1}{m}\sum_{i=1}^{m}\sqrt{\frac{1}{2}\sum_{j=1}^{n_i}(p_{ij}-q_{ij})^2}. \quad (5)$$

d_R is the average d_E across all loci standardized with the factor $\sqrt{1/2}$ to restrict the values to the interval [0,1]. It is one only if two OTUs are fixed for different alleles, but if one or both OTUs are not fixed and they have no alleles in common, d_R is not equal to one. d_R fulfills the distance properties (Nei et al., 1983), but it is not Euclidean. This follows from the identity $d_R = 1 - s_{SM}$ for homozygous inbred lines and the fact that $1 - s_{SM}$ is not Euclidean (Gower and Legendre, 1986).

Assuming that (i) F_1 was the cross between two homozygous inbred lines $P1$ and $P2$ and (ii) O was an inbred offspring derived from the F_1 cross, Melchinger et al. (1991) showed that d_R fulfilled following two genetical properties:

$$d_R(F_1, P1) = d_R(F_1, P2) = d_R(P1, P2)/2, \quad (6)$$

$$d_R(P1, O) + d_R(P2, O) = d_R(P1, P2). \quad (7)$$

The first property can be illustrated geometrically as three points F_1, $P1$ and $P2$ forming a line with F_1 lying in its center.

Based on these two properties, Melchinger et al. (1991) derived theoretical results that d_R estimates between two homozygous inbreds are linearly related to the coancestry coefficient (Malecot, 1948). Consequently, d_R is suitable for studying the relationship between the genetic dissimilarity of inbreds based on allelic informative marker data and the coefficient of coancestry (Malecot, 1948). This linear relationship is also desired in surveys (i) investigating the assembly and validation of core collections and the identification of duplicates in seed banks and (ii) uncovering pedigree relationships among OTUs as needed for the detection of essentially derived varieties in plant breeding.

Modified Rogers' Distance. Wright (1978) and Goodman and Stuber (1983) modified d_R by assigning each allele one dimension in the modified Rogers' distance (d_W):

$$d_W = \frac{1}{\sqrt{2m}}\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(p_{ij}-q_{ij})^2}. \quad (8)$$

Obviously, $d_W = 1/\sqrt{2m}d_E$ and as an Euclidean distance with values in [0,1] it can be used for the same applications as recommended for d_E. Like d_R, d_W is not equal to one in the case of multiple alleles, even if the two OTUs have no allele in common.

Consider two populations $\pi 1$ and $\pi 2$ in Hardy-Weinberg equilibrium and their hybrid population $\pi 1 \times \pi 2$. Based on results of Falconer and Mackay (1996), and assuming biallelism and absence of epistasis, Melchinger (1999) derived the following relationship between the mean of these populations:

$$\begin{aligned}\Delta H(\pi 1 \times \pi 2) &= \mu_{\pi 1 \times \pi 2} - (\mu_{\pi 1} + \mu_{\pi 2})/2 \\ &= \sum_i y_i^2 \delta_i = \\ &= \sum_i d_{W_i}^2(\pi 1, \pi 2)\delta_i,\end{aligned} \tag{9}$$

where ΔH is the panmictic-midparent heterosis (Lamkey and Edwards, 1999), δ_i is the dominance effect at QTL i, and y_i is the difference in gene frequencies. Consequently, a linear relationship between ΔH and d_W^2 is expected under the above conditions. Therefore, d_W^2 is especially suitable in studies based on allelic informative marker data for examining (i) the prediction of heterosis with genetic dissimilarities or (ii) the establishment of heterotic groups. Furthermore, d_W can be used for the same applications as suggested for d_E, owing to its Euclidean property.

Cavalli-Sforza and Edward's Chord Distance. Cavalli-Sforza and Edwards (1967) developed a genetic distance to analyze blood group allele frequencies in human populations. In this coefficient, an OTU with allele frequencies $p_1, p_2, \ldots, p_n$ is represented by the vector $(\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_n})$. Such a vector is always of unit length and, thus, the OTU is located on a surface of a hypersphere with radius one considering one locus. The distance between two OTUs is then directly proportional to the length of the chord connecting the points representing the OTUs. In particular, for two OTUs with no allele in common, d_{CE} is equal to one (Wright, 1978). For multiple loci the distances of all loci are combined by applying the Pythagorean theorem in many dimensions, so that the square of the distance between the OTUs is given by the sum of squared distances for each locus:

$$d_{CE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(1 - \sum_{j=1}^{n_i}\sqrt{p_{ij}q_{ij}})}. \tag{10}$$

d_{CE} ranges from zero to one even in the case of multiple alleles, which is an advantage over d_R and d_W. It can be shown that:

$$d_{CE} = \frac{1}{\sqrt{2m}}\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\sqrt{p_{ij}} - \sqrt{q_{ij}})^2}. \tag{11}$$

Thus, d_{CE} is similar to d_W except that it uses the square root of the allelic frequencies as coordinates and is consequently an Euclidean distance. d_{CE} was developed based on Kimura's (1954) model of "selective drift", by assuming that (i) the mutation rate is small and (ii) variation in selection pressure is rapid and haphazard. It seems doubtful that seed bank and plant breeding materials have evolved according to this model, because selection pressure is rather directed than rapid and haphazard. However, if allelic informative marker data are available and one can assume the "selective drift" model, then d_{CE} is a proper coefficient to investigate phylogenetic relationships among populations. Since d_{CE} is Euclidean, it can be used for the same tasks as proposed for d_E.

Reynolds' Dissimilarity. Reynolds et al. (1983) used the coancestry coefficient θ (Malecot, 1948) as the basis for a measure of genetic dissimilarity for short term evolution, when the divergence between populations with a common ancestral population may be regarded as being caused solely by drift:

$$d_{RE} = -ln(1 - \theta), \tag{12}$$

where

$$\theta = \frac{\sum_{i=1}^{m}\left(\frac{1}{2}\sum_{j=1}^{n_i}(p_{ij} - q_{ij})^2 - \frac{1}{2(2n-1)}\left[2 - \sum_{j=1}^{n_i}(p_{ij}^2 + q_{ij}^2)\right]\right)}{\sum_{i=1}^{m}\left(1 - \sum_{j=1} p_{ij}q_{ij}\right)}. \tag{13}$$

For populations completely fixed at each locus (i.e., two homozygous inbred lines) θ is equal to one and d_{RE} is undefined. Thus, d_{RE} is neither a distance nor Euclidean. d_{RE} was developed assuming that an ancestral population was split into several subpopulations of the same size, which subsequently diverged due to drift. In such a situation, d_{RE} is expected to increase linearly with the time since the populations diverged (Weir, 1996), i.e., $d_{RE} \approx t/2N$, where N is the subpopulation size and t the time measured in generations after divergence of the two populations. Thus, if mutation and selection can be neglected, and drift is the major evolutionary force, then d_{RE} is an appropriate dissimilarity coefficient for investigating the phylogenetic relationships among populations based on allelic informative marker data.

A recent application of d_{RE} was described by Labate et al. (2003), who examined relationships among U.S. maize landraces with SSR markers and assumed that an ancestral population split into several subpopulations diverging mainly due to drift. Mutation is known to have only small effects on genetic diversity compared with other forces and, thus, can safely be ignored in short term evolution scenarios. However, neglecting selection as an evolutionary force in plant breeding or in seed bank populations seems questionable in most instances.

Nei's Standard Genetic Dissimilarity. In contrast to d_{CE} and d_{RE}, where it is assumed that populations diverged due to random genetic drift, Nei (1972) suggested a dissimilarity coefficient based on mutation and drift, often referred to as Nei's standard dissimilarity. This measure is intended to estimate the average number of codon substitutions per locus and was defined as:

$$d_{N72} = -\ln\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} p_{ij}q_{ij}}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n_i} p_{ij}^2 \sum_{i=1}^{m}\sum_{j=1}^{n_i} q_{ij}^2}}. \tag{14}$$

Table 3 R^2 values obtained by a procrustes analysis with a published data set of seven CIMMYT maize populations (Reif et al., 2003) for seven dissimilarity coefficients based on differences in allele frequencies (Euclidean (d_E), Rogers' (d_R), modified Rogers' (d_W) and Cavalli-Sforza and Edward's (d_{CE}) distance and Reynolds' (d_{RE}), Nei's (1972) (d_{N72}), and Nei et al.'s (1983) (d_{N83}) dissimilarity coefficient).

	d_E	d_R	d_W	d_{CE}	d_{RE}	d_{N72}
d_R	0.0014					
d_W	0.0000	0.0014				
d_{CE}	0.0038	0.0047	0.0038			
d_{RE}	0.0592	0.0636	0.0592	0.0787		
d_{N72}	0.0307	0.0336	0.0307	0.0474	0.0103	
d_{N83}	0.0209	0.0233	0.0209	0.0228	0.0281	0.0172

Nei (1978) extended d_{N72} with a bias factor. If two OTUs differ in all alleles, d_{N72} is not defined, because it becomes $-\ln 0$. Thus, d_{N72} is neither a distance nor Euclidean. d_{N72} was developed based on the infinite-allele model (Kimura and Crow, 1964) assuming that an ancestral population split into various subpopulations, which diverged due to drift and mutation. If (i) the mutation-drift balance is maintained throughout the evolutionary process, (ii) selection is absent, and (iii) the dissimilarity is not very large, then $d_{N72} = 2vt$, where v is the mutation rate per locus and generation and t is the time measured in generations after divergence of the two populations (Nei et al., 1983). Under the above conditions, d_{N72} is suitable for investigating phylogenetic relationships among populations based on allelic informative marker data but otherwise, the same constrains apply as for d_{RE}.

Nei et al.'s (1983) Dissimilarity. Assuming the infinite allele model (Kimura and Crow, 1964), Nei et al. (1983) suggested in a simulation study a dissimilarity coefficient, which is quite efficient in recovering the true evolutionary tree when it is reconstructed from allele frequency data (Nei and Kumar, 2000):

$$d_{N83} = \frac{1}{m}\sum_{i=1}^{m}\left(1 - \sum_{j=1}^{n_i}\sqrt{p_{ij}q_{ij}}\right), \quad (15)$$

which equals d_{CE}^2. Nevertheless, the result of the simulation study depends heavily on the underlying evolutionary model of the simulation scenario. d_{N83} was not developed based on a specific genetic model and it is neither a distance nor Euclidean. Thus, application of d_{N83} in surveys for detecting phylogenetic relationships among populations seems questionable. For homozygous inbred lines, $d_{N83} = d_R$ and, hence, it could be used for the same applications as d_R.

Genetic Dissimilarity Coefficients for allelic non-informative marker data. With allelic non-informative marker data and two OTUs under consideration one can form a 2×2 table with entries v_{ij} (number of bands in common between both OTUs), w_{ij} (number of bands present in the ith OTU and absent in the jth OTU), x_{ij} (number of bands absent in the ith OTU and present in the jth OTU), and y_{ij} (number of bands absent from both OTUs).

The simple matching coefficient is one of the oldest similarity coefficients (Sneath and Sokal, 1973):

$$s_{SM} = \frac{v_{ij} + y_{ij}}{v_{ij} + w_{ij} + x_{ij} + y_{ij}}. \quad (16)$$

For homozygous inbred lines, $d_{SM} = 1 - s_{SM} = d_R$ and therefore, can be used for the same applications as suggested for d_R.

Jaccard (1908) suggested the similarity coefficient:

$$s_J = \frac{v_{ij}}{v_{ij} + w_{ij} + x_{ij}}. \quad (17)$$

The Dice coefficient (Dice, 1945) is defined as:

$$s_D = \frac{2v_{ij}}{2v_{ij} + w_{ij} + x_{ij}}. \quad (18)$$

The dissimilarity $d_D = 1 - s_D$ is also called the Nei-Li distance (Nei and Li, 1979) and is related to $d_J = 1 - s_J$ by a monotonic function.

In contrast to s_{SM}, both s_J and s_D do not involve negative matches (y_{ij}). For instance, if the probability of non-amplification of bands is high and absence of bands in both OTUs cannot be interpreted as a common characteristic, it is appropriate to apply coefficients s excluding negative matches (s_J and s_D).

In contrast to $1 - s$, $\sqrt{1-s}$ is a distance and Euclidean for all three presented coefficients (Gower and Legendre, 1986). Thus, they could be used to examine relationships among OTUs based on allelic non-informative marker data in combination with multivariate methods requiring dissimilarity coefficients with the Euclidean property.

Relationships Among Dissimilarity and Similarity Coefficients

If (i) band absence or presence can be interpreted as two alleles of one locus and (ii) the OTUs under consideration are homozygous inbreds, then the following relationships exist between the s and d coefficients:

$$d_R = d_W^2 = d_{N83} = d_{CE}^2 = \frac{1}{2m}d_E^2 = 1 - s_{SM}. \quad (19)$$

Gower (1975) proposed a method of comparing different multivariate analyses of the same data set, also known as Procrustes analysis (Cox and Cox, 2000). We used this approach to il-

Table 4 Overview of the genetical and mathematical properties of dissimilarity coefficients based on allelic informative marker data: Euclidean (d_E), Rogers' (d_R), modified Rogers' (d_W), and Cavalli-Sforza and Edward's distance (d_{CE}) and Nei's (1972) (d_{N72}), Nei et al.'s (1983) (d_{N83}), and Reynolds' (d_{RE}) dissimilarity coefficient.

Dissimilarity coefficient	Properties
d_E	No underlying genetic concept. Suited to investigate relationships among operational taxonomic units (OTUs) with multivariate methods that require Euclidean distances (principal coordinate analysis, hierarchical cluster analysis, classification, hierachical classification, and graph theory).
d_R	Linearly related to coefficient of coancestry. Appropriate to examine (i) the assembly and validation of core collections and (ii) the uncovering of pedigree relationships among OTUs such as the detection of essentially derived varieties in plant breeding or the identification of duplicates and collection gaps in seed banks.
d_W	${d_W}^2$ is linearly related to panmictic-midparent heterosis. Therefore, d_W^2 is appropriate to examine (i) the prediction of heterosis with genetic distances or (ii) the establishment of heterotic groups.
d_{CE}	Based on Kimura's (1954) model of "selective drift". If one can assume the "selective drift" model, then d_{CE} is a proper coefficient to investigate the phylogenetic relationships among populations.
d_{RE}	Based on a model that an ancestral population splits into several subpopulations of the same size, which diverge due to drift. Thus, if mutation and selection can be neglected and drift is the major evolutionary force, then d_{RE} is suitable for investigating the phylogenetic relationships among populations.
d_{N72}	Based on the infinite-allele model (Kimura and Crow, 1964). If one can assume the infinite-allele model, then d_{N72} is suitable for investigating phylogenetic relationships among populations.
d_{N83}	For homozygous inbred lines, $d_{N83} = d_R$ and, hence, d_{N83} is also linearly related to the coancestry coefficient (Malecot, 1948). Therefore, d_{N83} can be used for inbred lines for the same applications as d_R.

lustrate the differences between the dissimilarity coefficients based on allele frequency differences (Table 1).

The Procrustes analysis is based on the pairwise comparison between two sets of dissimilarities d_{ij} and d^*_{ij} $(i, j = 1, 2, \ldots, n)$ among the same sample of n OTUs. Rather than concentrating on the distances themselves, geometric points $P_i (i = 1, \ldots, n)$ of the n OTUs are constructed to give rise to all the inter-distances d_{ij}. The coordinates of these points were obtained with Kruskal's non-metric multidimensional scaling (Cox and Cox, 2000). Kruskal's non-metric multidimensional scaling is a technique to represent OTUs in a reduced space while preserving the distance relationships among them with high fidelity. It is not limited to Euclidean distance matrices and can produce ordinations of objects from any dissimilarity matrix. Similarly, the coordinates of the points $P^*_i (i = 1, \ldots, n)$ are found for the dissimilarities d^*_{ij} by applying again Kruskal's non-metric multidimensional scaling. The two configurations are then matched for best fit by means of translation, rotation, and reflection. The criterion of best fit adopted is the minimization of the 'residual' sum of squares $R^2 = \sum_{i=1}^{n} {d_E}^2(P_i, P^*_i)$, where $d_E(P_i, P^*_i)$ is the Euclidean distance between corresponding points P_i and P^*_i.

We compared the seven coefficients d based on allele frequency differences (Table 1) of a published data set of seven tropical CIMMYT maize populations (Reif et al. 2003) by subjecting them pairwise to the procrustes analysis. The resulting R^2 matrix (Table 3) was then used as input for Kruskal's non-metric multidimensional scaling (Figure 1). The same analyses were also performed with other data sets and yielded similar results (data not shown). All analyses were performed with Version 2 of the Plabsim software (Frisch et al. 2000), which is implemented as an extension to the statistical software R (Ihaka and Gentleman 1996).

The distance between d_E and d_W is zero (Table 3), because $d_W = \sqrt{2m} d_E$. Both measures clustered together with d_R and d_{CE} (Figure 1). This is in accordance with the expectations, because (i) d_{CE} equals d_W except that the square roots of the allele frequencies are used as coordinates and (ii) d_R is the average d_E across all loci standardized by the factor $\sqrt{2}$. d_{N83} was positioned between d_E, d_W, d_R and d_{CE} on one side and d_{N72} and d_{RE} on the other side. This is not surprising because d_{N72} and d_{RE} are based on similar assumptions: an ancestral population split into subpopulations diverging by drift (d_{RE}) or by mutation and drift (d_{N72}). Both coefficients include an estimate of the allele frequencies of the ancestral population in contrast to the other measures. Consequently, our results indicate that the analogy of d_{N72} and d_{RE} in estimating the allele frequencies of the ancestral population has a stronger influence on the property of the coefficients than the choice of the evolutionary model assuming drift and mutation or only drift. Summarizing, some coefficients are mathematically related or were developed assuming similar evolutionary models.

Conclusions

Our investigations show that genetical (Table 4) and mathematical (Tables 1 and 2) properties of dissimilarity measures are of crucial importance when choosing a genetic dissimilarity coefficient for analyzing molecular marker data. The presented results can assist experimenters in the choice of dissimilarity measures that allow the extraction of the maximum amount of

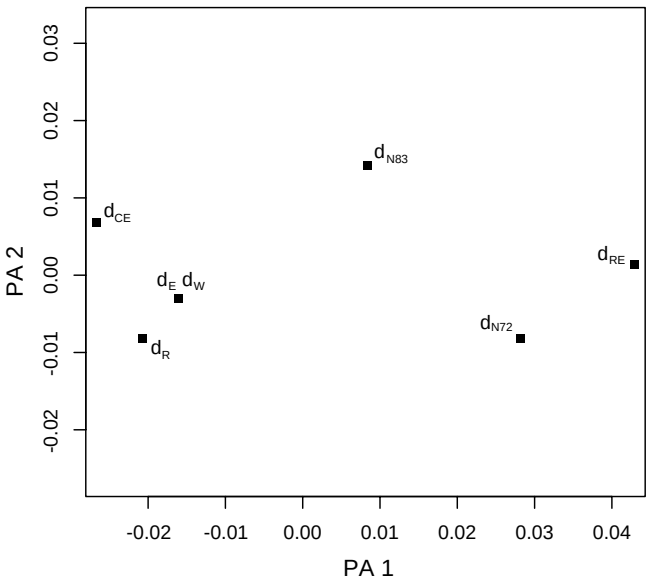

Figure 1 First two principal axes (PA) of Kruskal's non-metric multidimensional scaling for comparison of Euclidean (d_E), Rogers' (d_R), modified Rogers' (d_W) and Cavalli-Sforza and Edward's distance (d_{CE}) and Nei's (1972) (d_{N72}), Nei et al.'s (1983) (d_{N83}), and Reynolds' (d_{RE}) dissimilarity coefficient based on a procrustes analysis with a published data set of seven CIMMYT maize populations (Reif et al., 2003).

information from genetic data for given objectives. Thus, they facilitate the interpretation of findings from molecular marker studies on a theoretically sound basis.

References

Brummer, E.C. 1999. Capturing heterosis in forage crop cultivar development. Crop Sci. 39:943–954.

Cavalli-Sforza, L.L., and A.W.F. Edwards. 1967. Phylogenetic analysis: models and estimation procedures. Am. J. Hum. Genet. 19:233–257.

Cox, T.F., and M.A.A. Cox, 2000. Multidimensional Scaling. Chapman and Hall, Florida.

Dice, L.R. 1945. Measures of the amount of ecologic association between species. Ecology 26:297–302.

Dubreuil, P., and A. Charcosset. 1998. Genetic diversity within and among maize populations: a comparison between isozyme and nuclear RFLP loci. Theor. Appl. Genet. 96:577–587.

Falconer, D.S., and T.F.C. Mackay. 1996. Introduction to Quantitative Genetics. 4th eds Longman Group Ltd, London.

Frisch M., M. Bohn, A.E. Melchinger. 2000. Plabsim: Software for simulation of marker-assisted backcrossing. J. of Heredity 91:86-87.

Geerlings, H., A.J. Van Oeveren, J.E. Pot, R.C. van Schaik. 1999. AFLP-Quantar^{TM} Pro Image analysis software. http://www.keygene-products.com.

Goodman, M.M. 1972. Distance analysis in biology. Syst. Zool. 174–186.

Goodman, M.M., and C.W. Stuber. 1983. Races of maize: VI. Isozyme variation among races of maize in Bolivia. Maydica 28:169–187.

Gower, J.C. 1975. Generalised Procrustes analysis. Psychometrika 40:33–50.

Gower, J.C. 1985. Measures of similarity, dissimilarity and distances. p. 397–405. *In* S. Kotz, N.L. Johnson, and C.B. Read (eds) Encyclopedia of Statistical Sciences, Vol. 5. Wiley, Ney York.

Gower, J.C., and P. Legendre. 1986. Metric and Euclidean properties of dissimilarity coefficients. J. Classification 3:5–48.

Hamrick, J.L., and M.J.W. Godt. 1990. Allozyme diversity in plant species. p. 43–63. *In* A.H.D. Brown, M.T. Clegg, A.L. Kahler, and B.S. Weir (eds.) Plant Population Genetics, Breeding and Genetic Resources. Sinauer, Sunderland, MA.

Ihaka, R., and R. Gentleman. 1996. A language for data analysis and graphics. J. of Computational and Graphical Statistics, Vol. 5. 3:299–314.

Jaccard, P. 1908. Nouvelles recherches sur la distribution florale. Bull. Soc. Vaud. Sci. Nat. 44:223–270.

Jansen, R.C., H. Geerlings, A.J. van Oeveren, and R.C. van Schaik. 2001. A comment on codominant scoring of AFLP markers. Genetics 158:925–926.

Kimura, M. 1954. Process leading to quasi-fixation of genes in natural populations due to random fluctuation of selection intensities. Genetics 39:280–295.

Kimura, M., and J.F. Crow. 1964. The number of alleles that can be maintained in a finite population. Genetics 49:725–738.

Labate, J.A., K.R. Lamkey, S.H. Mitchell, S. Kresovich, H. Sullivan, and J.S.C. Smith. 2003. Molecular and historical aspects of Corn Belt dent diversity. Crop Sci. 43:80–91.

Lamkey, K.R., and J.W. Edwards. 1999. Quantitative genetics of heterosis. Chapter 10. *In* J.G. Coors and S. Pandey (eds.) The Genetics and Exploitation of Heterosis in Crops. CSSA, Madison, WI.

Lombard, V., C.P. Baril, P. Dubreuil, F. Blouet, and D. Zhang. 2000. Genetic relationships and fingerprinting of rapeseed cultivars by AFLP: Consequences for varietal registration. Crop Sci. 40: 1417-1425.

Malecot, G. 1948. Les Mathématiques de l'Hérédite. Masson et Cie, Paris.

Melchinger, A.E. 1999. Genetic diversity and heterosis. Chapter 10. *In* J.G. Coors and S. Pandey (eds.) The Genetics and Exploitation of Heterosis in Crops. CSSA, Madison, WI.

Melchinger, A.E., M.M. Messmer, M. Lee, W.L. Woodman, and K.R. Lamkey. 1991. Diversity and relationships among U.S. maize inbreds revealed by restriction fragment length polymorphisms. Crop Sci. 31:669–678.

Messick, S.M., and R.P. Abelson. 1956. The additive constant problem in multidimensional scaling. Psychometrica 21:1–15.

Mohammadi, S.A., and B.M. Prasanna. 2003. Analysis of genetic diversity in crop plants – salient statistical tools and considerations. Crop Sci. 43:1235–1248.

Nei, M. 1972. Genetic distance between populations. Am. Nat. 106:283–292.

Nei, M. 1978. The theory of genetic distance and evolution of human races. Jpn. J. Hum. Genet. 23:341–369.

Nei, M., and S. Kumar. Molecular Evolution and Phylogenetics. 2000. Oxford Univ. Press, New York.

Nei, M., F. Tajima, and Y. Tateno. 1983. Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. J. Mol. Evol. 19:153–170.

Nei, M., and W.H. Li. 1979. Mathematical models for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76:5269:5273.

Piepho, H.-P., and G. Koch. 2000. Codominant analysis of banding data from a dominant marker system by normal mixtures. Genetics 155:1459–1468.

Reynolds, J., B.S. Weir, C.C. Cockerham. 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105:767–779.

Reif, J.C., A.E. Melchinger, X.C. Xia, M.L. Warburton, D.A. Hoisington, S.K. Vasal, G. Srinivasan, M. Bohn, and M. Frisch. 2003. Genetic distance based on sim-

ple sequence repeats and heterosis in tropical maize populations. Crop Sci. 43:1275–1282.

Rogers, J.S. 1972. Measures of genetic similarity and genetic distance. p. 145–153. Studies in Genetics VII. University of Texas Publication 7213, Austin, Texas.

Smith, J.S.C., O.S. Smith, S.L. Bowen, R.A. Tenborg, and S.J. Wall. 1991. The description and assessment of distances between inbred lines of maize. III. A revised scheme for the testing of distinctiveness between inbred lines utilizing DNA RFLPs. Maydica 36:213–226.

Sneath, P.H.A., and R.R. Sokal. 1973. Numerical Taxonomy. Freeman, San Francisco, CA.

Treuren, R. van, L.J.M. van Soest, and Th.J.L. van Hintum. 2001. Marker-assisted rationalisation of genetic resources collections: a case study in flax using AFLPs. Theor. Appl. Genet. 103:144–152.

Weir, B.S. 1996. Genetic Data Analysis II. p. 91. 2nd ed. Sinauer Associates. Inc., Sunderland. M.A.

Wright, S. 1978. Evolution and Genetics of Populations, Vol. IV. p. 91. The Univ. of Chicago Press, Chicago, IL.

Wheat genetic diversity trends during domestication and breeding

Jochen C. Reif*, Pingzhi Zhang†, Susanne Dreisigacker*, Marilyn L. Warburton†, Maarten van Ginkel†, David Hoisington†, Martin Bohn*‡, and Albrecht E. Melchinger*

** Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany, † International Maize and Wheat Improvement Center (CIMMYT), Apartado Postal 6-641, 06600 Mexico, D.F., Mexico, ‡ University of Illinois, Urbana, IL 61801, USA*

Abstract

It has been claimed that plant breeding reduces genetic diversity in elite germplasm, which could seriously jeopardize the continued ability to improve crops. The main objective of this study was to examine the loss of genetic diversity in bread wheat during (i) its domestication, (ii) the change from traditional landrace cultivars (LC) to modern breeding varieties, and (iii) 50 years of international breeding. We studied 253 CIMMYT or CIMMYT-related modern wheat cultivars, LC, and *Triticum tauschii* accessions with 90 simple sequence repeat (SSR) markers covering the entire wheat genome. A loss of genetic diversity was observed from *T. tauschii* to LC, and from LC to the elite breeding germplasm. Wheat's genetic diversity was narrowed from 1950 to 1989, but was enhanced from 1990 to 1997. Our results indicate that breeders averted the narrowing of the wheat germplasm base and subsequently increased the genetic diversity through the introgression of novel materials. The LC and *T. tauschii* contain numerous unique alleles that were absent in modern wheat cultivars. Consequently, both LC and *T. tauschii* represent useful sources for broadening the genetic base of elite wheat breeding germplasm.

DOMESTICATION and modern plant breeding have presumably narrowed the genetic base of bread wheat (*Triticum aestivum*), which could jeopardize future crop improvement. Tetraploid varieties of domesticated wheat were derived from a single tetraploid progenitor, *T. dicoccoides*, the donor of the A and B genomes (1). Soon after the domestication of *T. dicoccoides*, free-threshing forms evolved from less amenable hulled genotypes, known as *T. turgidum*. Wheat species with tetraploid genomes were subsequently involved in a fateful experiment: accidental crosses with the wild diploid species *T. tauschii*, the donor of the D genome. This gave rise to hexaploid wheat *T. aestivum*, also known as bread wheat (2, 3, 4). The number of independent crosses between the progenitors of *T. aestivum* is considered limited (5, 6), resulting presumably in a loss of diversity.

Through the centuries, mutation generated new alleles, while recombination created novel allele combinations. The genetic variation available in all this was subsequently reduced by genetic drift and selection, both natural and that of early farmers, which eventually resulted in landrace cultivars (LC) adapted to specific conditions of their habitats. During the past century so many of the traditional LC were continually replaced by modern wheat cultivars (MWC) that currently only about 3% of the wheat-growing area is sown with LC (7). The MWC were bred with a limited number of LC in their pedigree and it is postulated that MWC contain less genetic diversity than LC (8).

A popular hypothesis is that an extended period of plant breeding and intensive selection have further reduced genetic diversity among cultivars, narrowing the germplasm base available for future breeding advances (9). Cultivation of germplasm with a narrow genetic base entails a risk due to genetic vulnerability. This risk is that mutations in pest populations or changes in environmental conditions may bring about stresses that the cultivar could not cope with, and therefore could lead to severe crop losses. This risk was brought sharply into focus in 1970 with the outbreak of the southern corn leaf blight (10). The first signs that germplasm with a narrow genetic base might also lead to disasters in wheat came from several severe epidemics of shoot fly (*Atherigona* spp.) and karnal bunt (*Tilletia indica*) in India in the 1970s (11). Nevertheless, plant breeding does not inevitably lead to a loss of genetic diversity. Reduction in diversity caused by intensive selection can be counterbalanced by introgression of novel germplasm.

During the last 40 years, the International Maize and Wheat Improvement Center (CIMMYT) has had a tremendous impact on spring wheat. In all developing countries, excluding China, approximately 86% of the spring bread wheat area in 1997 was sown with CIMMYT or CIMMYT-related germplasm involving at least one CIMMYT ancestor (7). CIMMYT's wheat germplasm is therefore exceptionally suitable for investigation whether breeding has reduced genetic diversity in wheat in a detrimental manner.

Examining the loss of genetic diversity in bread wheat during (i) its domestication, (ii) the change from traditional LC to modern breeding varieties, and (iii) 50 years of international breeding requires molecular analyses that incorporate comprehensive samples of MWC, LC, and their progenitors. In this article, we report the first such extensive molecular diversity analysis of wheat, which used a sample of 253 MWC, LC, and *T. tauschii* accessions and 90 simple sequence repeat markers (SSR) that provide a broad coverage of the wheat genome.

Corresponding author: Albrecht E. Melchinger, Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany. Email: melchinger@uni-hohenheim.de. Received March 28, 2004.

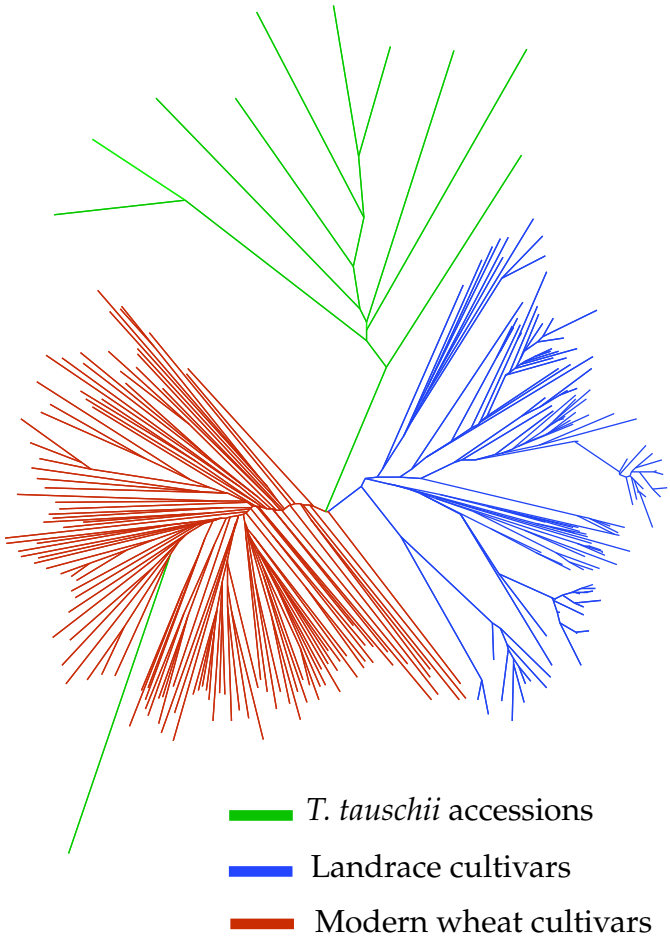

Figure 1 Fitch-Margoliash tree for the 11 *T. tauschii* accessions, 119 wheat landrace cultivars, and 123 modern wheat cultivars using the Rogers' distance (RD).

Material and Methods

Plant Materials. We have chosen 123 CIMMYT, CIMMYT-related, and other MWC according to their total area sown, year of release, contribution to the development of new important lines (key parents), and geographic distribution, taken from an impact study (7). The MWC were divided into five time periods according to the year of release: Period 1: 1950-1965, Period 2: 1966-1973, Period 3: 1974-1981, Period 4: 1982-1989, and Period 5: 1990-1997. Each period included a minimum of 20 MWC. Detailed information on the 123 MWC is available as supporting information at *http://www.cimmyt.org/english/webp/support/publications/support_materials/ssr_mw1.htm*. Five Mexican and four Turkish spring wheat LC composed of three to 25 sub-lines were added to our study, resulting in a total of 119 LC genotypes. Detailed information about the LC is published elsewhere (12). Additionally, 11 *T. tauschii* accessions were chosen for analysis, six collected in Iran, two in China, and three of unknown origin (detailed information is published elsewhere (13)).

Simple Sequence Repeat Genotyping. The plants were genotyped by the Applied Biotechnology Center at CIMMYT. Details of the protocol are published elsewhere (12). Briefly, DNA was extracted by the CTAB method and the SSR regions were amplified by PCR with flourescent-labeled primers. PCR products were size-separated on an ABI Prism 377 DNA Sequencer (Perkin Elmer Biotechnologies, Foster City, CA) and classified to specific alleles by GENESCAN and GENOTYPER software programs (12). MWC were genotyped with a set of 90 SSRs (51 EST and 39 genomic derived), covering the entire wheat genome. The LC were fingerprinted with a subset of the SSRs consisting of 41 markers (14). The SSR information was obtained from IPK (Gatersleben, Germany) and DuPont (Wilmington, DE). In addition, the SSR markers Taglgap, Taglut (15), and WMC56 developed by the Wheat Microsatellite Consortium (Agrogene, France) were used. The *T. tauschii* SSR genotypes were obtained with 28 SSRs mapping to the D genome, as described elsewhere (13). Details for all SSRs are given in Table B at *http://www.cimmyt.org/english/webp/support/publications/support_materials/ssr_mw1.htm*.

Statistics. Rogers' genetic distance (RD) (16) was estimated among pairs of genotypes, considering the absence of an SSR marker band as a missing value. Based on RD estimates, the Fitch-Margoliash least-squares algorithm implemented in the computer program Phylip was used to construct a phylogenetic tree (17). Standardized numbers of alleles per locus (N_a) were estimated by re-sampling nine plants per group (MWC, LC, and *T. tauschii*) (18). Gene diversity (H) was calculated for

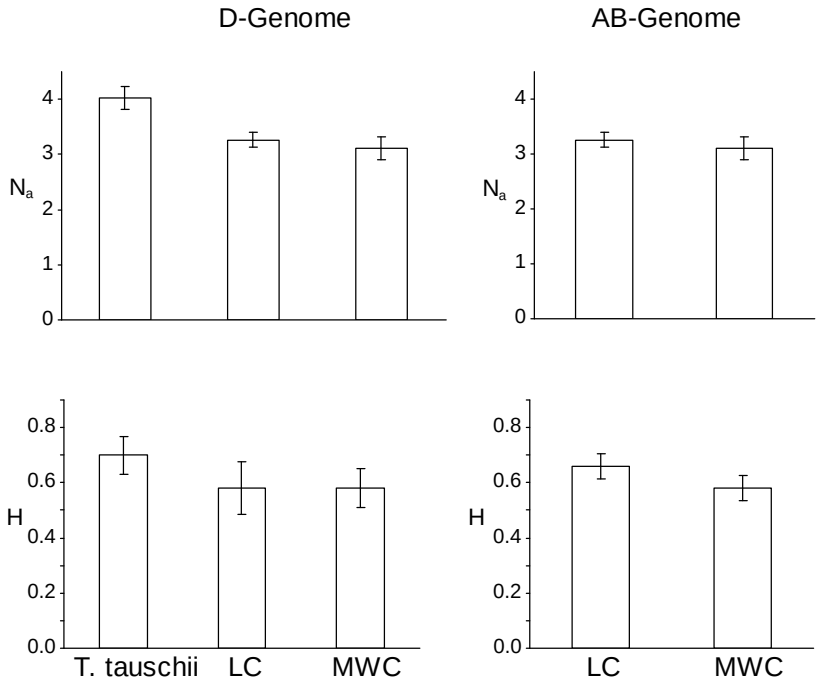

Figure 2 Standardized number of alleles per locus (N_a) and gene diversity (H) of 11 *T. tauschii* accessions, 119 landrace cultivar (LC), and 123 modern wheat cultivar (MWC) genotypes. Values for the D genome are based on 14 SSRs, and for the AB genomes on 27 SSRs.

MWC, LC, and *T. tauschii* (19). Standard errors of N_a and H were determined by a bootstrap procedure over SSRs. Relative loss of gene diversity between two germplasm groups was calculated as: $\Delta H = 1-(H_1/H_2)$, where H_1 and H_2 denote the gene diversities of the two germplasm groups (20). Furthermore, the average number of unique alleles per SSR was determined for comparison between *T. tauschii* and LC, as well as between LC and MWC. The coefficient of parentage (COP) (21) was estimated among pairs of MWC as described elsewhere (22).

A linear relationship between COP and RD is expected under certain simplified assumptions (23). Pearson's correlation coefficient (r) was calculated between RD estimates based on SSRs and COP values based on all pairs of MWC with COP $\geq$ 0.05. Trends of RD- or COP-based genetic diversity estimates of the MWC during the last 50 years of wheat breeding were examined by performing multiple regression analyses of these measures on the time periods, following established procedures (24). All analyses were performed with Version 2 of the Plabsim software (25), which is implemented as an extension to the statistical software R (26).

Results and Discussion

Relationships among Modern Wheat Cultivars, Landraces, and *T. tauschii*. The Fitch-Margoliash phylogenetic tree of all 253 genotypes revealed a clear separation of *T. tauschii* accessions from LC and MWC with only one *T. tauschii* accessions positioned in the group of MWC (Fig. 1). This result reflects the long isolation of the two gene pools after domestication, as well as the divergence caused by selection, drift, and mutation. LC and MWC also formed two separated main clusters. This outcome can be explained by (1) the limited number of LC used as the germplasm base for the development of MWC and (2) selection and drift during the breeding of MWC. The LC and *T. tauschii* accessions are quite diverse from MWC, indicating their potential as a source of novel germplasm for wheat breeding.

Loss of Diversity from *T. tauschii* to Landraces. We observed a non-significant decrease in N_a and H from *T. tauschii* accessions to LC (Fig. 2) but a significant ($P < 0.1$) relative diversity loss ($\Delta H = 0.19$). These results, together with the findings of 2.5 unique alleles per locus present in *T. tauschii* but not in LC, indicate a reduction in genetic variation during the process of wheat domestication. This is in agreement with previous studies reporting that the *T. tauschii* genome contains considerably more genetic variation than the D genome of hexaploid wheat (27, 28). The reduction in genetic diversity is probably the product of the relatively young history of the wheat crop, the presumably small founder population, and the intensive long-term selection for agronomic traits. Thus, the initial steps of crop domestication probably caused a severe population bottleneck.

Loss of Diversity from Landraces to Modern Wheat Cultivars. No change in H from LC to MWC was observed for the D genome, but H decreased slightly from LC to MWC for the AB genomes (Fig. 2). Combining all SSRs, a relative loss of gene diversity ΔH of 0.05 was revealed from LC to MWC. Together with the observation that 1.9 unique alleles per locus were present in LC but absent in MWC, this indicated a substantial genetic diversity loss from LC to MWC. Possible explanations are those already stated in discussing the clustering pattern of

Table 1 Average Rogers' distance (above diagonal) and coefficient of parentage (below diagonal) for 123 CIMMYT and CIMMYT-related wheat cultivars grouped into five time periods (Period 1: 1950-1965, Period 2: 1966-1973, Period 3: 1974-1981, Period 4: 1982-1989, and Period 5: 1990-1997). The mean of the standard errors of Rogers' distances is 0.053.

Period	1	2	3	4	5
1		0.477	0.487	0.481	0.487
2	0.109		0.463	0.462	0.474
3	0.103	0.144		0.443	0.465
4	0.089	0.122	0.141		0.445
5	0.094	0.136	0.150	0.155	

the material in our study. The loss of genetic diversity may indicate an elimination of undesired or even deleterious alleles, but may also reflect an erosion of alleles valuable for plant improvement and future demands of producers and consumers. The latter hypothesis was supported by various surveys reporting the potential of LC as a source of novel useful allelic variation (29, 30).

Loss of Diversity during 50 Years of International Breeding. The global impact of the wheat breeding program of CIMMYT has been significant and well documented (31). The main objectives of this breeding program were high and stable yields across mega-environments combined with widely effective disease resistances. An average annual increase in yield of 0.88% was reported from 1960 to 1990 (32). Nevertheless, there has been growing public concern that the tremendous enhancements of yield by modern breeding would go hand in hand with a large decrease in diversity (33), which could threaten future selection progress.

The average RD and COP between MWC of different time periods (Table 1) showed that the relatedness of germplasm increased with decreasing differences in time periods, indicating the presence of drift and/or selection. The average RD and COP between MWC of adjacent time periods revealed that the relatedness of the germplasm decreased with increasing time periods. This reflects the effects of second-cycle breeding, where the next breeding cycle is generated by intermating the best genotypes of the previous cycle.

Pairwise RD within a period regressed on the period number corroborated a significant ($P < 0.05$) quadratic trend. This indicates a narrowing of genetic diversity among major CIMMYT MWC from Period 1 to Period 4, but an enhancement from Period 4 to Period 5 (Fig. 3A)

COP determines the similarity between two individuals using the concept of identity by descent. Pairwise 1-COP values between individuals within time periods present therefore an alternative measure of genetic diversity. Although the correlation (r) between RD and 1-COP across all 7503 data points was only 0.48 ($P < 0.01$), we observed also a significant ($P < 0.05$) quadratic trend between SSR-based RDs and time periods, with an increase in diversity for the last time period studied (Fig. 3B). The low correlation between RD and 1-COP can be explained by several simplifying assumptions underlying the COP estimation including (1) unrelated founder individuals, (2) equal parental genome distribution, and (3) the absence of selection, mutation or drift (34).

Owing to the length of a breeding cycle and the low multiplication rate of wheat after the initial cross is made, it takes approximately 10-12 years for a newly developed cultivar to reach the market and influence the genetic diversity on a large scale. Therefore, the decreasing pairwise 1-COP values and pairwise RDs from Period 1 to Period 4 reflect the reduction in genetic diversity until the late 1970s. This reduction in diversity levels might be explained by the ´´Early Green Revolution" (35), which was characterized by breeding semi-dwarf varieties possessing a higher yielding potential due to an increased harvest index and better lodging tolerance, especially under high fertilizer and water inputs. These high yielding new semi-dwarf MWC were based on a limited number of key parents and dominated rapidly the wheat germplasm base (11).

The increase in genetic diversity (1-COP values and RD) from Period 4 to Period 5 can be explained by a change in the breeding strategy of CIMMYT in the late 1970s. CIMMYT's wheat breeding program aimed at increasing genetic diversity on a large scale by taking into account the need for biological diversification, environmental sustainability, and durable resistance to combat ever-evolving pathogens. In parallel, remained the wide geographic adaptation of the germplasm an important breeding goal (36, 37). The breeding germplasm was broadened with (i) spring and winter wheat from different regions, (ii) exotic germplasm such as Chinese or Brazilian wheat cultivars, (iii) LC from many regions, and (iv) wild relatives such as *Agropyrum* derivatives (31, 36).

Our results indicate that CIMMYT breeders successfully increased the genetic diversity through introgression of various novel wheat materials once they realized the danger of narrowing down their germplasm base. Grain yield of spring bread wheat has been systematically increased through genetic improvement from Period 4 to Period 5 (32) without reducing the genetic diversity (Fig. 3B). Thus, the enhancement of yield in plant breeding does not necessarily cause a loss of genetic diversity.

Sources of Novel Genetic Variation for Wheat Breeding. Over the last 100 years, the development and successful application of wheat breeding has produced high-yielding MWC on which current agriculture is based. Yet, ironically, it is the plant breeding process itself that threatens the genetic base upon which breeding depends. A report commissioned by the National Academy of Sciences, in response to the 1970 southern corn leaf blight disaster, recommended placing more emphasis on collecting and preserving the genetic diversity in crop species (10). One result of that report was the foundation of germplasm banks such as the one at CIMMYT, where approximately 150 000 accessions of wheat and its wild relatives are conserved.

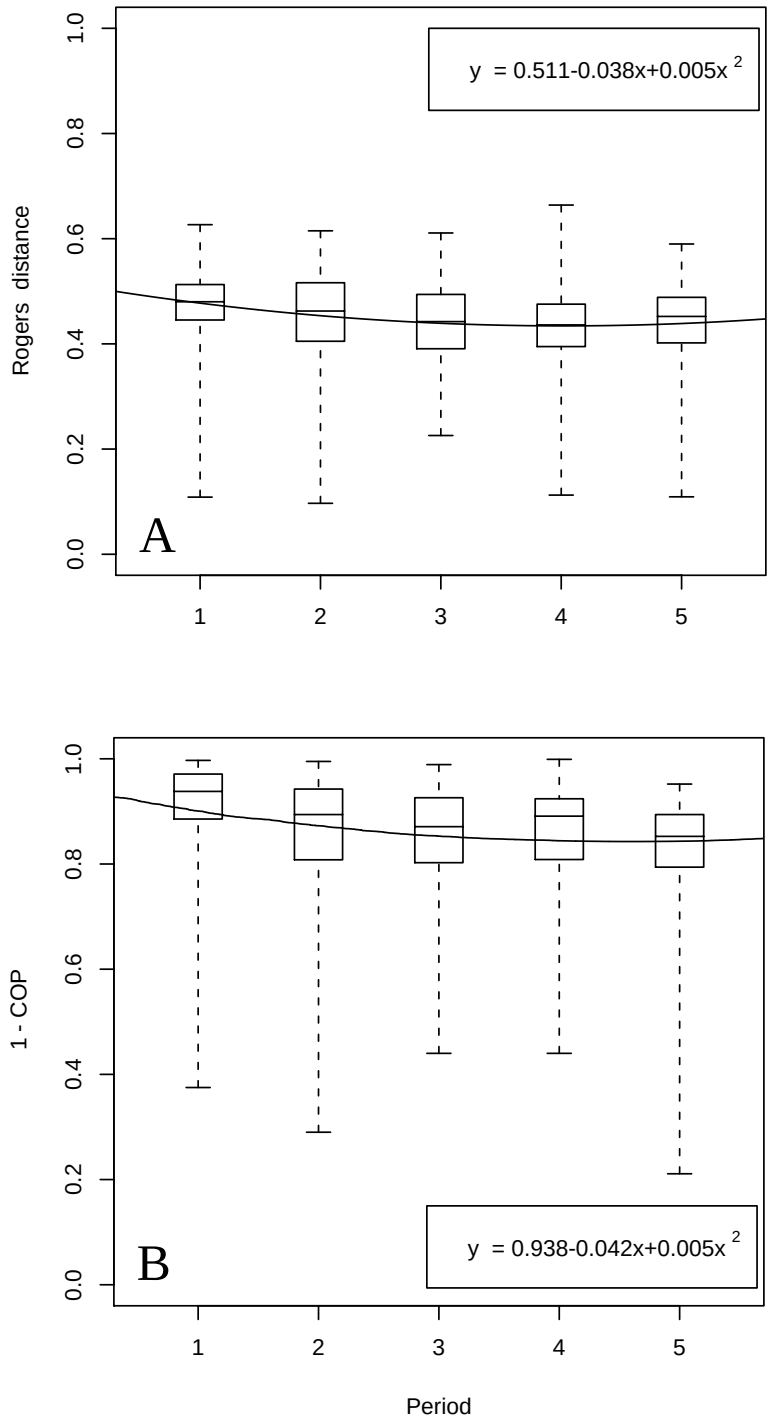

Figure 3 Boxplot of pairwise (A) Rogers' distances and (B) 1-COP values for 123 CIMMYT and CIMMYT-related wheat cultivars grouped into five time periods. The closed boxes comprise values between the 25% and 75% quantiles.

A classical way experienced scientists and research staff use to identify useful novel genes and alleles in genetic resources is to look for potentially useful traits. This may happen during routine maintenance and systematic screening of collections, or as a spin-off of pre-breeding and breeding programs carried out for other purposes. Once a desired trait has been identified, backcrossing can be used to introduce it into elite breeding germplasm. This approach works well when the trait of interest is controlled by one or a few genes but many traits important to agriculture, such as yield, show polygenic inheritance.

The identification of genes of agronomic importance therefore requires more sophisticated methods (20). After their localization in the genome, a systematic search of novel alleles can be conducted in genetic resources via new approaches such as association mapping (38). New valuable genetic variants can then be introgressed systematically, applying marker-assisted backcrossing or genetic transformation. Consequently, the genetic potential present in genetic resources can be unlocked, facilitating a sustainable future selection gain in plant breeding.

Acknowledgment

We thank Bent Skovmand for providing the wheat material for this study. We are also indebted to the Vater & Sohn Eiselen-Stiftung, Ulm, and the German Federal Ministry of Economic Co-operation and Development, for their grateful financial support and collaboration within the project "Efficient management of genetic diversity in wheat: DNA marker for use in wheat breeding programs and gene banks". We dedicate this article to Dr. Norman Borlaug, the father of the "Green Revolution".

References

1. Kimber, G. & Feldman, M. (2001), *Wild Wheat. An Introduction* (Special Report 353, College of Agriculture, Univ. of Missouri – Colombia), pp. 129–131.
2. Kihara, H. (1944) *Agric. Hortic.* 19, 13–14.
3. McFadden, E.S. & Sears, E.R. (1946) *Journal of Heredity* 37, 81–89.
4. Salamini, F., Özkan, H., Brandolini, A., Schäfer-Pregl. R. & Martin, W. (2002) *Nat. Rev. Genet.* 3, 429–441.
5. Dvorák, J., Luo, M.C., Yang, Z.L. & Zhang, H.B. (1998) *Theor. Appl. Genet.* 67, 657–670.
6. Talbert, L.E., Smith, L.Y. & Blake, M.K. (1998) *Genome* 41, 402–407.
7. Smale, M., Reynolds, M.P., Warburton, M., Skovmand, B., Trethowan, R., Singh, R.P., Ortiz-Monasterio, I. & Crossa, J. (2002) *Crop Sci.* 42, 1766–1779.
8. Frankel, O.H. (1970) *World Agric.* 19, 9–14.
9. Tanksley, S.D. & McCouch, R. (1997) *Science* 277, 1063–1066.
10. Anonymous (1972) *Genetic Vulnerability of Major Crops* (National Academy of Sciences, Washington, DC).
11. Dalrymple, D.G. (1986) in *Development and Spread of High-yielding Wheat Varieties in Developing Countries 7th ed.* (US Agency for International Development, Washington DC).
12. Dreisigacker, S., Zhang, P., Warburton, M., Skovmand, B., Hoisington, D. & Melchinger, A.E. (2004) *Crop Sci.* 44,381–388.
13. Zhang, P., Dreisigacker, S., Melchinger, A.E., van Ginkel, M., Hoisington, D. & Warburton, M.L. (2004) *Molecular Breeding* in review.
14. Dreisigacker, S., Zhang, P., Warburton, M., Skovmand, B., Hoisington , D. & Melchinger, A.E. (2004) *Crop Sci.* in press.
15. Devos, K.M., Bryan, G.J., Collins, A.J., Stephenson, P. & Gale, M.D. (1995) *Theor. Appl. Genet.* 100, 247–252.
16. Rogers, J.S. (1972) *Studies in Genetics VII* (Univ. of Texas Publication, Austin, Texas) pp. 145–153.
17. Felsenstein, J. (1993) *PHYLIP – Phylopgeny Inference Package, Version 3.5c* (Department of Genetics, University of Washington, Seattle).
18. Reif, J.C., Xia, X.C., Melchinger, A.E., Warburton, M.L., Hoisington, D., Beck, D., Bohn, M. & Frisch, M. (2004) *Crop Sci.* 44, 326–334.
19. Nei, M. (1987) *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York).
20. Vigouroux, Y., McMullen, M., Hittinger, C.T., Houchins, K., Schulz, L., Kresovich, S., Matsuoka, Y. & Doebley, J. (2002) *Proc. Natl. Acad. Sci. USA* 99, 9650–9655.
21. Malecot, G. (1948) *Les mathématiques de l'hérédite* (Masson et Cie, Paris).
22. Martin, S.K.St. (1982) *Crop Sci.* 22, 151–152.
23. Melchinger, A.E., Messmer, M.M., Lee, M., Woodman, W.L. & Lamkey, K.R. (1991) *Crop Sci.* 31, 669–678.
24. Snedecor, G.W. & Cochran, W.G. (1980) *Statistical Methods* (Iowa State University Press, Ames).
25. Frisch, M., Bohn, M. & Melchinger, A.E. (2000) *Journal of Heredity* 91, 86–87.
26. Ihaka, R. & Gentleman, R. (1996) *Journal of Computational and Graphical Statistics* 3, 299–314.
27. Lubbers, E.L., Gill, K.S., Cox, T.S. & Gill, B.S. (1991) *Genome* 34, 354–361.
28. Lelley, T., Stachel, M., Grausgruber, H. & Vollmann, J. (2000) *Genome* 43, 661-668.
29. Cox, T.S., Wilson, W.J., Gill, D.L., Leath, S., Bockus, W.W. & Browder, L.E. (1992) *Plant Disease* 76, 1061–1064.
30. Villareal, R.L., Davila, G.F. & Kazi, A.M. (1995) *Cereal Research Communications* 23, 127–132.
31. Rajaram, S. (1994) *Wheat Breeding at CIMMYT: Commemorating 50 Years of Research in Mexico for Global Wheat Improvement* (Wheat special report No 29. CIMMYT, Mexico D.F., Mexico).
32. Sayre, K.D., Rajaram, S. & Fischer, R.A. (1997) *Crop Sci.* 37, 36–42.
33. Harlan, J.R. (1972) *Journal of Environmental Quality* 1, 212-215.
34. Cox, T.S., Murphy, J.P. Rodgers, D.M. (1986) *Proc. Natl. Acad. Sci. USA* 83, 5583–5586.
35. Evenson, R.E. & Gollin, D. (2003) *Science* 300, 758–762.
36. Reeves, T., Rajaram, S., van Ginkel, M., Trethowan, R., Braun, H. & Cassaday, K. (1999) *New Wheats for a Secure, Sustainable Future* (CIMMYT, Mexico D.F., Mexico).
37. Rajaram, S. & van Ginkel, M. (2001) in *The World Wheat Book, A History of Wheat Breading*, eds. Bonjean, A.P. & Angus, W.J. (Lavoisier Publishing, Paris, France), pp. 579–610.
38. Lynch, M. & Walsh, B. (1997) *Genetics and analysis of quantitative traits* (Sinauer Assoc., Sunderland, MA), pp. 413.

Genetic Diversity Determined within and among CIMMYT Maize Populations of Tropical, Subtropical, and Temperate Germplasm by SSR Markers

J. C. Reif, X. C. Xia, A. E. Melchinger,* M. L. Warburton, D. A. Hoisington, D. Beck, M. Bohn, and M. Frisch

ABSTRACT

Genetic diversity in maize (*Zea mays* L.) plays a key role for future breeding progress. The main objectives of our study were to (i) investigate the genetic diversity within and among CIMMYT maize populations by simple sequence repeat (SSR) markers, (ii) examine genotype frequencies for deviations from Hardy-Weinberg equilibrium (HWE) at individual loci, and (iii) test for linkage disequilibrium (LD) between pairs of loci. Twenty-three maize populations and pools established in 1974, which mostly comprise germplasm from different racial complexes adapted to tropical, subtropical intermediate-maturity, subtropical early-maturity, and temperate megaenvironments (ME), were fingerprinted by 83 SSR markers covering the entire maize genome. Across all populations, 27% of the SSR markers deviated significantly from HWE with an excess of homozygosity in 99% of the cases. We observed no evidence for genome-wide LD among pairs of loci within each of the seven tropical populations analyzed. Estimates of genetic differentiation (G_{ST}) between populations within MEs averaged 0.09 and revealed that most of the molecular variation was found within the populations. Principal coordinate analysis based on allele frequencies of the populations revealed that populations adapted to the same ME clustered together and, thus, supported clearly the ME structure.

Genetic diversity in maize is a valuable natural resource and plays a key role for future breeding progress. Germplasm collections as a source of genetic diversity must be well characterized for efficient management and effective exploitation. Achieving this goal in curating of gene banks is hampered by rising costs, decreasing budgets, and large collection sizes. The germplasm collection sizes should be optimized to provide maximal preservation of genetic variation and minimal redundancy with regard to genotypes, gene complexes, or possibly even genes (Kresovich et al., 1992).

Association mapping was proposed as one approach to detect genes and alleles of interest in germplasm collections (Lynch and Walsh, 1997). The resolution of association studies in a sample depends on the extent of linkage disequilibrium (LD) across the genome. LD (or the correlation between alleles of different loci) depends generally on the genealogy of the germplasm. Besides this, drift and selection within populations can also cause LD. The genomic structure of LD must be empirically determined before embarking on association studies because it can vary among samples of germplasm. The advent of PCR-based molecular markers such as SSRs has created an opportunity for fine-scale genetic characterization of germplasm collections. Since SSR markers are highly polymorphic (Smith et al., 1997), easy to generate, and highly repeatable (Heckenberger et al., 2002), they can be used for large-scale investigations as needed in the case of genetic resources (Powell et al., 1996).

CIMMYT developed and improved from 1964 until 1973 a wide array of maize germplasm. Populations were established with materials from a single racial complex. In 1974, a major shift in the organization of the germplasm was initiated. Germplasm from different racial complexes was mixed and more than 100 populations were established to (i) reduce the large collection of germplasm from CIMMYT's gene bank to a number that can be handled efficiently in a breeding program and (ii) use the combining ability of different germplasm sources for intrapopulation improvement. In addition, 30 broad-based back-up pools were formed as an insurance against narrowing the germplasm base of the populations (CIMMYT, 1998). These pools and populations have played an important role in maize breeding and production in developing countries and have been exploited as sources of new germplasm for temperate regions (Ron Parra and Hallauer, 1997). Detailed knowledge about LD and genetic diversity of these populations would increase the efficiency of their use in breeding. However, little is known about the molecular diversity in tropical and subtropical maize populations (Warburton et al., 2002) and information about LD in this germplasm is entirely lacking.

The main objectives of our study were to characterize the population genetic structure of 23 CIMMYT maize populations as a basis for an efficient use of this germplasm in breeding programs. In particular we (i) investigated the molecular genetic diversity within and among 23 of CIMMYT's maize populations, (ii) examined genotype frequencies for deviations from Hardy-Weinberg equilibrium at individual loci, and (iii) tested for LD between pairs of loci.

J.C. Reif, A.E. Melchinger, and M. Frisch, Institute of Plant Breeding, Seed Science, and Population Genetics, Univ. of Hohenheim, 70593 Stuttgart, Germany; M. Bohn, Crop Science Dep., Univ. of Illinois, 1102 South Goodwin Avenue, Urbana, IL 61801; M.L. Warburton, D.A. Hoisington, and D. Beck, International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641 06600 Mexico D.F., Mexico; X.C. Xia, Institute of Crop Breeding and Cultivation, Chinese Academy of Agric. Sciences, Zhongguancun South Street 12, 100081, Beijing, China. 30 April 2003. *Corresponding author (melchinger@uni-hohenheim.de).

Published in Crop Sci. 44:326–334 (2004).

677 S. Segoe Rd., Madison, WI 53711 USA

Abbreviations: CIMMYT, International Maize and Wheat Improvement Center; HWE, Hardy-Weinberg equilibrium; ME, megaenvironment; LD, linkage disequilibrium; MRD, modified Roger's distance; PC, principal coordinate; PCoA, principal coordinate analysis; QTL, quantitative trait locus; SSR, simple sequence repeat.

MATERIALS AND METHODS

Genetic Material

Twenty-three populations and pools (further referred to as populations, if not stated otherwise) of CIMMYT's maize program were investigated including tropical, subtropical, and temperate adapted material. The germplasm was grouped on the basis of its adaptation to four megaenvironments (MEs): tropical, subtropical intermediate-maturity, subtropical early-maturity, and temperate (Table 1). The populations were maintained periodically by recombining a minimum of 420 plants per population each generation for several decades.

SSR Analyses

Forty-eight individuals from each of the seven tropical populations and 21 individuals from each of the 16 subtropical and temperate populations were chosen at random and analyzed individually. DNA was extracted employing a modified CTAB procedure (Saghai-Maroof al., 1984). We used the set of 83 SSR markers described by Warburton et al. (2002), which provides uniform coverage of the entire maize genome. Primers and PCR conditions were described in detail by Warburton et al. (2002). SSRs were multiplexed for a maximum efficiency. Fragments were separated using acrylamide gels run on an ABI 377 automatic DNA sequencer. Fragment sizes were calculated with GeneScan 3.1 (Perkin Elmer/Applied Biosystems, Foster City, CA) using the Local Southern sizing method (Elder and Southern, 1987); allele identity was assigned by Genotyper 2.1 software (Perkin Elmer/Applied Biosystems) and the two inbred lines, CML51 and CML292, as controls. Map position for all SSRs were based on the Pioneer Composite 1999 linkage map obtained from the MaizeDB website (http://nucleus.agron.missouri.edu/).

Statistical Analyses

The number of alleles per locus (further referred to as allelic richness) was determined for the entire set of 672 individuals analyzed and for various subsets within this collection (populations, MEs). The existence of population- or ME-specific alleles was determined. The total gene diversity (H_T) based on SSR data across all populations was decomposed into (i) gene diversity between individuals within each population (H_S) and (ii) gene diversity between populations within each ME (H_{ME}) according to Nei (1987, p. 164) and Chakraborty (1980). Confidence intervals for H_S values were obtained by a bootstrap procedure with resampling across markers and individuals. The coefficient of gene differentiation (G_{ST}) was used as a measure of genetic differentiation between populations of the same ME and different MEs and was calculated according to Nei (1987, p. 190). G_{ST} is the proportion of the total genetic diversity that is due to differences between MEs. The fixation index F_{IS} for each population was estimated according to Nei (1987, p. 164) as one minus the observed heterozygosity divided by the expected heterozygosity.

The average number of alleles, the number of unique alleles, H_S, and H_{ME} depend on the number of individuals analyzed per population. In the tropical populations, 48 individuals were sampled, whereas in the subtropical and temperate populations only 21 individuals were sampled per population. Therefore, we used a resampling strategy to obtain comparable estimates: a random sample of 21 individuals from each of the tropical populations was chosen for the above analyses,

Table 1. Description of the 23 CIMMYT maize populations analyzed in this study, grouped according to the adaptation to the four megaenvironments (MEs).

Population	Cycle	Germplasm description
		Tropical ME
P21	5	Composed of seven Tuxpeño races plus some families from P124.
P22	6	Includes Tuxpeño and ETO Blanco germplasm as well as germplasm from Central America.
P25	0	Composed of white flint selections from crosses among germplasm from Mexico, Columbia, the Caribbean, Central America, India, Thailand and the Philippines.
P29	5	Broad genetic base including Tuxpeño, Cuban flints, and ETO.
P32	5	On the basis of white flint germplasm from South America, Cuba, Mexico and the U.S. Corn Belt.
P43	5	Tuxpeño synthetic composed of 16 S_1 lines.
P124	21	Mainly based on Tuxpeño germplasm, but includes also some materials from Central America, the Carribean, and Zaire.
		Subtropical intermediate-maturity ME
P33	2	Contains predominantly Argentinian (Cateto) flints.
P34	5	Includes Cuban flints, ETO, Tuxpeño, and germplasm from the U.S. Corn Belt, India, and Nepal.
P42	4	An advanced generation of ETO selected for short-plant type and crossed with Illinois Corn Belt material.
P45	3	Includes U.S. Corn Belt germplasm, Tuxpeño, Cuban flints, Puerto Rico composite, and collections from the Dominican Republic.
P47	2	Consists largely of Tuxpeño germplasm plus some U.S. Corn Belt lines.
P131	14	A broad based pool including white flint segregates from Ecuador, Argentina, India, Mexico, Pool 32, and Pool 33, but also germplasm from Mexico, U.S. Corn Belt, Brazil, Uruguay, Argentina, China, Pakistan, Yugoslavia, Lebanon, Guatemala, Venezuela, Peru, Cuba, and the Dominican Republic.
P134	20	Includes germplasm from the Mexican lowlands and highlands, the U.S. Corn Belt, southern USA, Puerto Rico, Pakistan, Hungary, China, Peru, Pakistan, Lebanon, Nicaragua, and Guatemala.
		Subtropical early-maturity ME
P46	1	Represents a superior fraction (240 half-sib families) of Pool 29, which is based on germplasm from Europe, Lebanon, the USA Corn Belt, China, Indonesia, and South America.
P48	5	Composed of U.S. Corn Belt germplasm, southern European germplasm, and 54 half-sib families from Pool 30.
P127	20	Includes germplasm from the USA, China, Lebanon, Pakistan, and several European countries.
P128	14	On the basis of crosses between white dent segregates from P127 and Hungarian germplasm from Pool 30 and various other germplasm.
P130	15	Composed of germplasm from Europe, China, Lebanon, Mexico, South America, and the U.S. Corn Belt.
		Temperate ME
P139	12	Contains germplasm from the tropical lowlands and highlands, subtropical, and temperate areas.
P140	12	On the basis of germplasm from Europe.
P141	12	Includes predominantly U.S. Corn Belt germplasm plus germplasm from China, Korea, and Lebanon.
P142	12	On the basis of germplasm from Mexico, Peru, Bolivia, Pakistan, Hungary, USA, and Yemen.

sampling was repeated 1000 times, and the results were averaged.

The modified Roger's distance (MRD) between two populations or individuals was calculated according to Wright (1978, p. 91) and Goodman and Stuber (1983). Standard errors of MRD estimates were calculated by a bootstrap procedure with resampling across markers and individuals. Associations among operational taxonomic units were revealed by principal coordinate analysis (PCoA) (Gower, 1966) based on MRD values. PCoA were performed for (i) the 23 populations and (ii) all individuals of the populations from each ME. In the latter case, individuals with more than 30% missing values were excluded. All analyses were performed with Version 2 of the Plabsim software (Frisch et al., 2000), which is implemented as an extension to the statistical software R (Ihaka and Gentleman, 1996).

Alleles with frequencies smaller than 0.10 were pooled for each locus for tests of Hardy-Weinberg equilibrium (HWE) and LD because disequilibrium coefficients have large variances with rare alleles. The population genetic software Arlequin (Schneider et al., 2000) was used for tests of HWE at individual loci and LD between pairs of loci. Software Arlequin uses the procedure described by Guo and Thompson (1992) to detect significant departures from HWE. LD between all pairs of loci was tested within each of the seven tropical populations using a likelihood-ratio test, whose empirical distribution is obtained by a permutation procedure (Slatkin and Excoffier, 1996). This test assumes HWE at each locus and, thus, only loci with no significant deviation from HWE were included in the analysis. An LD analysis of the 16 subtropical and temperate populations was not considered due to the small sample size of 21 individuals per population. In testing for both HWE and LD, the Bonferroni correction for multiple tests was applied (Snedecor and Cochran, 1980).

RESULTS

Population Genetic Analyses

We found a total of 666 alleles for the 83 SSR markers in the 672 genotypes (Table 2). All 83 marker loci analyzed were polymorphic across all 672 individuals. The average proportion of missing data across genotypes was 8%. At the population level, an average of 76 loci was polymorphic with a maximum of 81 (Pl42) and a minimum of 63 (Pl40). The percentage of loci with

Table 2. Genetic diversity within and among 23 maize populations as revealed by SSRs.

Population	No. of individuals	Avg. no. of alleles per locus	No. of unique alleles†	Gene diversity‡	Fixation index (F_{IS})	95% confidence interval for F_{IS}
			Tropical germplasm			
P21	48	4.20	13	0.51^{ab}	0.27	(0.19,034)
P22	48	4.14	4	0.51^{ab}	0.25	(0.17,0.31)
P25	48	3.93	8	0.51^{ab}	0.23	(0.15,0.29)
P29	48	4.24	7	0.55^{b}	0.25	(0.17,0.32)
P32	48	3.47	3	0.45^{a}	0.31	(0.19,0.35)
P43	48	3.55	3	0.48^{ab}	0.28	(0.19,0.35)
P124	48	4.22	1	0.53^{ab}	0.22	(0.13,0.30)
Total	*336*	*6.07*	*86*	*0.56*		
			Resampled tropical germplasm§			
P21	21	3.74	12	0.50^{ab}	0.27	(0.18,0.35)
P22	21	3.71	6	0.50^{ab}	0.24	(0.15,0.31)
P25	21	3.56	7	0.50^{ab}	0.22	(0.15,0.30)
P29	21	3.83	5	0.54^{ab}	0.23	(0.15,0.32)
P32	21	3.08	2	0.45^{ab}	0.27	(0.19,0.35)
P43	21	3.23	3	0.47^{ab}	0.28	(0.19,0.38)
P124	21	3.79	1	0.52^{ab}	0.21	(0.11,0.30)
Total	*147*	*5.62*	*35*	*0.55*		
			Subtropical intermediate-maturity germplasm			
P33	21	3.73	4	0.55^{ab}	0.44	(0.30,0.50)
P34	21	3.71	3	0.52^{ab}	0.41	(0.28,0.47)
P42	21	3.58	2	0.54^{ab}	0.36	(0.24,0.43)
P45	21	3.80	4	0.57^{b}	0.37	(0.24,0.42)
P47	21	3.65	7	0.54^{ab}	0.37	(0.24,0.42)
P131	21	3.78	1	0.59^{ab}	0.65	(0.54,0.72)
P134	21	3.90	5	0.56^{ab}	0.41	(0.28,0.48)
Total	*147*	*5.86*	*37*	*0.62*		
			Subtropcal early-maturity germplasm			
P46	21	3.48	3	0.54^{ab}	0.41	(0.27,0.46)
P48	21	3.36	1	0.51^{ab}	0.36	(0.26,0.42)
P127	21	3.78	8	0.55^{ab}	0.38	(0.25,0.44)
P128	21	4.00	4	0.58^{b}	0.43	(0.29,0.48)
P130	21	3.75	3	0.55^{ab}	0.40	(0.27,0.47)
Total	*105*	*5.43*	*23*	*0.60*		
			Temperate germplasm			
P139	21	4.13	6	0.58^{b}	0.40	(0.28,0.43)
P140	21	3.81	4	0.55^{ab}	0.41	(0.29,0.45)
P141	21	3.69	5	0.57^{b}	0.39	(0.26,0.43)
P142	21	3.96	5	0.57^{b}	0.42	(0.30,0.47)
Total	*84*	*5.34*	*22*	*0.61*		
Grand total	*672*	*8.02*		*0.62*		

† Number of unique alleles with respect to the total number of 672 alleles found in all 23 populations.
‡ Gene diversity values followed by the same letters are not different at the 0.05 significance level according to a bootstrap procedure.
§ A random sample of 21 individuals from each tropical population was chosen. Sampling was repeated 1000 times and the results averaged.

significant ($P < 0.01$) deviations from HWE varied from 14% (P48) to 40% (Pl34) with an average of 27% (Fig. 1). In 99% of the cases, deviations from HWE were attributable to an excess of homozygosity. F_{IS} values ranged from 0.21 (Pl24) to 0.65 (Pl31) (Table 2) across germplasm types. The number of unique alleles for the various MEs ranged from 22 (temperate ME) to 86 (tropical ME). Significant differences ($P < 0.05$) were found among the H_S values within the 23 populations with a range from 0.45 (P32) to 0.59 (Pl31) and a mean of 0.54. Gene diversity between populations within a given ME (H_{ME}) was highest in the subtropical intermediate-maturity populations (0.62) and smallest in the tropical populations (0.56). The coefficient of gene differentiation G_{ST} between the populations within MEs averaged 0.09 with little variation among the four MEs. Gene differentiation G_{ST} between the MEs was 0.02.

The number of LD tests performed and significant test results were (1275, 3) for P21, (1891, 2) for P22, (1275, 3) for P25, (1653, 6) for P29, (1035, 13) for P32, (946, 1) for P43, and (1485, 6) for Pl24. Thus, the proportion of significant two-locus LD tests was below the number of expected false positives with an experimentwise error rate of $\alpha = 0.05$.

Relationships between Populations

Values of MRD between pairs of populations averaged 0.28 and ranged from 0.20 (P22 × Pl24) to 0.41 (P32 × P48) with significant differences ($P < 0.01$) between MRD estimates (Table 3). The average MRD between all pairs of populations within MEs ranged from 0.22 (temperate ME) to 0.26 (subtropical intermediate-maturity ME) and averaged 0.25. The average MRD between all pairs of populations of different MEs was maximum for tropical × subtropical early-maturity populations (0.32) and minimum for subtropical early-maturity × temperate populations (0.24).

In the PCoA based on MRD estimates of all populations, the first two principal coordinates (PC) explained a total of 34.2% of the molecular variance (Fig. 2). The tropical populations were separated from the other

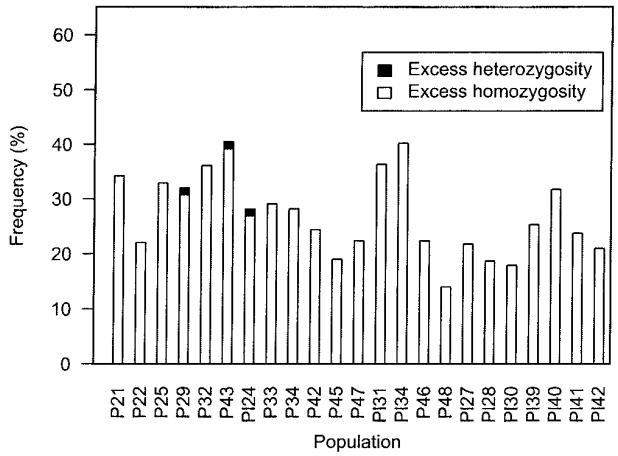

Fig. 1. Frequency of loci with significant ($P \leq 0.01$) Hardy-Weinberg equilibrium tests in the 23 CIMMYT maize populations.

Table 3. Modified Roger's distances between populations (average standard error 0.02).

	Tropical ME						Subtropical intermediate maturity ME							Subtropical early maturity ME						Temperate ME			
Population	P21	P22	P25	P29	P32	P43	Pl24	P33	P34	P42	P45	P47	Pl31	Pl34	P46	P48	P127	P128	P130	P139	P140	P141	P142
P21		0.22	0.27	0.23	0.30	0.27	0.22	0.29	0.31	0.31	0.32	0.31	0.31	0.29	0.30	0.37	0.29	0.29	0.32	0.30	0.32	0.32	0.32
P22			0.25	0.23	0.28	0.26	0.20	0.27	0.30	0.31	0.30	0.30	0.29	0.27	0.31	0.36	0.29	0.29	0.33	0.29	0.32	0.31	0.31
P25				0.25	0.26	0.26	0.22	0.27	0.29	0.30	0.29	0.28	0.30	0.29	0.31	0.37	0.29	0.30	0.32	0.29	0.33	0.31	0.31
P29					0.28	0.26	0.21	0.27	0.30	0.29	0.30	0.29	0.29	0.27	0.29	0.36	0.28	0.29	0.32	0.30	0.31	0.31	0.30
P32			$\overline{\text{MRD}} = 0.25$			0.31	0.27	0.30	0.28	0.28	0.32	0.31	0.33	0.30	0.33	0.41	0.33	0.32	0.35	0.33	0.35	0.34	0.33
P43							0.23	0.29	0.33	0.33	0.33	0.32	0.33	0.32	0.33	0.39	0.31	0.32	0.34	0.32	0.33	0.33	0.35
P124								0.25	0.29	0.30	0.29	0.27	0.29	0.26	0.28	0.35	0.27	0.28	0.31	0.28	0.31	0.30	0.30
P33									0.24	0.26	0.24	0.26	0.25	0.24	0.25	0.33	0.24	0.26	0.27	0.23	0.27	0.26	0.25
P34										0.24	0.29	0.27	0.27	0.28	0.32	0.37	0.29	0.30	0.32	0.28	0.33	0.31	0.31
P42											0.28	0.28	0.27	0.27	0.29	0.34	0.28	0.28	0.30	0.27	0.31	0.28	0.29
P45												0.27	0.24	0.22	0.24	0.29	0.24	0.23	0.24	0.23	0.27	0.23	0.22
P47					$\overline{\text{MRD}} = 0.30$				$\overline{\text{MRD}} = 0.26$				0.26	0.26	0.28	0.36	0.27	0.28	0.31	0.28	0.30	0.29	0.29
P131														0.27	0.26	0.32	0.25	0.26	0.29	0.26	0.29	0.28	0.27
P134															0.24	0.34	0.24	0.23	0.27	0.23	0.28	0.26	0.26
P46																0.29	0.23	0.23	0.25	0.23	0.25	0.25	0.24
P48																	0.30	0.26	0.22	0.28	0.25	0.25	0.26
P127					$\overline{\text{MRD}} = 0.32$					$\overline{\text{MRD}} = 0.28$							$\overline{\text{MRD}} = 0.25$		0.27	0.23	0.25	0.25	0.24
P128																		0.22	0.22	0.21	0.21	0.23	0.22
P130																				0.23	0.21	0.22	0.23
P139																					0.22	0.21	0.21
P140																						0.21	0.22
P141																							0.21
P142			$\overline{\text{MRD}} = 0.31$								$\overline{\text{MRD}} = 0.27$						$\overline{\text{MRD}} = 0.24$			$\overline{\text{MRD}} = 0.21$			

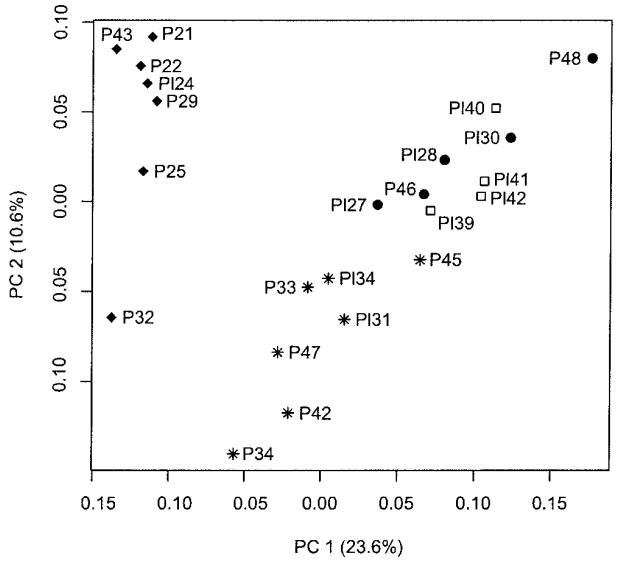

Fig. 2. Principal coordinate analysis of the 23 maize populations based on modified Roger's distance calculated from the allele frequencies of the populations. PC1 and PC2 are the first and second principal coordinates, respectively. Tropical (◆), subtropical intermediate-maturity (*), subtropical early-maturity (●), and temperate germplasm (□).

three ME populations by PC1. The subtropical intermediate-maturity populations were separated from the subtropical early-maturity and temperate populations by PC2. The temperate populations were positioned adjacent to the subtropical early-maturity populations.

In the tropical populations, individuals of P21, P22, P29, and PI24 clustered together and were clearly separated from individuals of P25, P32, and P43 (Fig. 3). Within the subtropical intermediate-maturity populations, individuals from P34 and P42 clustered together. Individuals of PI31 were widely spread across the two PCs, whereas individuals of P45 and P47 formed largely separated groups. In the subtropical early-maturity populations, PC1 clearly separated individuals of PI27 and P46 from individuals of PI30 and P48, whereas individuals of PI28 were positioned in between these two groups. The individuals of the four temperate pools were widely scattered across both PCs, with PI39 being most clearly separated from PI42.

DISCUSSION

Research on genetic diversity in maize with molecular markers has mostly concentrated on temperate inbred lines and their pedigree relationships as well as assignment to heterotic groups (Melchinger, 1999). Only a few studies have investigated the genetic diversity and structure of traditional maize populations from Europe (Gauthier et al., 2002) and the U.S. Corn Belt (Labate et al., 2003). In contrast to these populations, which have originated from different geographic regions and maintained separately by farmers and early breeders, the 100 populations and 30 pools at CIMMYT have a fairly short evolutionary history. These germplasm groups were created by breeders in 1974 mostly by intermating different racial complexes (Vasal et al., 1999). The intermixing of diverse germplasm within populations complicates detection of relationships among these populations based on pedigree information. Hence, we employed molecular markers to analyze associations among the CIMMYT maize pools and populations.

Hardy-Weinberg Equilibrium

The Hardy-Weinberg law describes the fundamental observation that in a large random-mating population both gene frequencies and genotype frequencies are constant across generations assuming absence of migration, mutation, and selection. The genotype frequencies are determined by the gene frequencies (Falconer and Mackay, 1996, p. 5). CIMMYT breeders have maintained their populations by planting a minimum of 20 rows and 21 plants per row. All plants consistent with the varietal description were shoot bagged and pollen from 10 rows was bulked to pollinate plants in the other 10 rows and vice versa. A minimum of 300 to 350 typical ears from the pollinated plants was chosen to represent each population. Considering the procedure to maintain the germplasm, it was expected that the populations would be in HWE after one generation of random mating. However, all 23 maize populations deviated significantly from HWE (Fig. 1) and showed a deficit of heterozygous loci (Table 2). This is in agreement with previous reports on other maize populations. Labate et al. (2000) investigated two random-mated maize populations and found that 27% of tests for deviation from Hardy-Weinberg equilibrium were significant, with deviations occurring due to an excess of homozygosity of 72 and 87%. Dubreuil and Charcosset (1998) also detected an excess of homozygosity in 10 populations from Europe and the U.S. using RFLP markers. In 17 open-pollinated populations assayed at 13 enzyme marker loci, 27% of Hardy-Weinberg tests were significant, with 94% showing an excess of homozygosity (Kahler et al., 1986).

The inbreeding of the populations observed in our study can be related to various causes: (i) positive assortative matings between individuals (homogamy), (ii) artificial subgrouping of individuals from populations, (iii) selection favoring homozygotes, and (iv) experimental errors during the laboratory assay for SSRs. Even though precautions were taken to avoid positive assortative mating between individuals, it cannot be excluded entirely because late flowering plants are preferentially crossed to late ones, and early flowering plants with early ones. However, only SSRs closely linked to QTL for flowering time should show a higher degree of homozygosity than expected under HWE. Assortative mating can be one reason for an artificial subgrouping of individuals, but a closer examination of the PCoAs (Fig. 3) did not provide any clue that the deficit of heterozygous individuals could be related to a subgrouping of individuals from populations. Selection favoring homozygotes is unlikely in maize, where fitness increases with heterozygosity. The choice of SSRs with tri- and higher repeats in our study and the use of the Local Southern sizing method to estimate the fragment

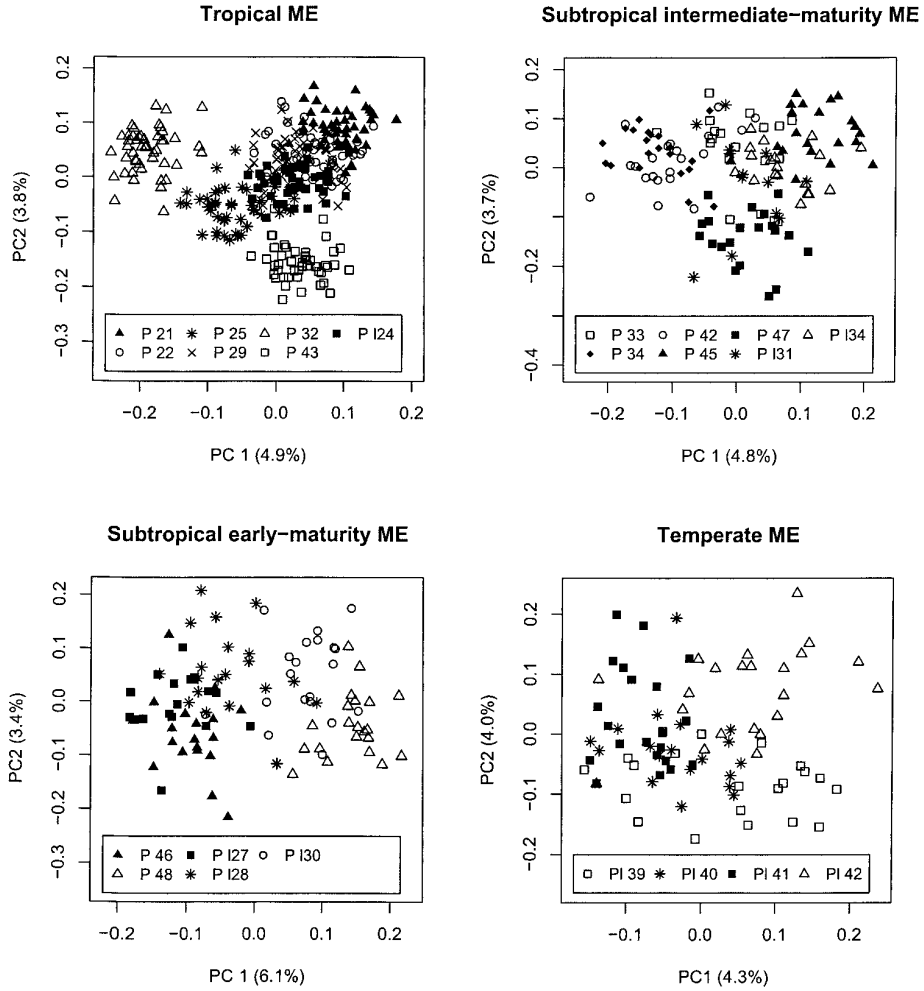

Fig. 3. Principal coordinate analyses of the individuals of the 23 populations grouped into four megaenvironments (ME) based on modified Roger's distance. PC1 and PC2 are the first and second principal coordinate, respectively, and numbers in parentheses refer to the proportion of variance explained by the principal component in the specific ME group.

sizes, which represents a conservative allele-calling procedure, reduce the laboratory error sources, which cause overestimation of heterozygosity. However, most experimental errors would lead to an overestimation of homozygotes because (i) a heterozygous locus carrying a null allele would be scored as a homozygous locus, (ii) alleles could not be detected because of competition during the PCR reaction, and (iii) the setting of the threshold of band intensity to detect alleles can be too strict.

Thus, experimental errors are probably the major cause of heterozygote deficiency within the populations apart from genuine genetic causes. To separate both sources, it would be prudent in future studies to include, besides the two inbred checks, their hybrid as a control to estimate the error rate for misscoring of heterozygous loci.

Linkage Disequilibrium

LD can result from and be maintained by epistasis (Falconer and Mackay, 1996, p. 16). It can also arise from admixture of populations with different gene frequencies, or from drift in small populations. Since population admixture happened during the establishment of the tropical populations recently, this could have caused LD. However, we found that less than 0.3% of the two-locus disequilibrium tests were significant, which can be explained by type I error alone. This is in accordance with a study reported by Stuber et al. (1980), who evaluated LD among eight enzyme loci in four long-term maize selection experiments. In contrast to these results, Remington et al. (2001) reported evidence of genome-wide LD among 47 SSRs for 102 maize inbred lines from temperate and tropical regions. LD was reduced but not eliminated by grouping lines into three empirically determined subpopulations. Nevertheless, artificial population admixture within the subpopulations caused by sampling lines from different germplasm sources could be one reason for the detection of LD in this survey. On one hand, the lack of LD in our study can be explained by the low-density marker map and the decrease of LD with successive generations of intermating since the establishment of the populations.

On the other hand, the sample size of 48 individuals per population, the precision in estimating haplotype frequencies with the EM algorithm (Excoffier and Slatkin, 1995), and the elimination of loci deviating from HWE (Fig. 1) result in a low power to detect LD. Further investigations are required to examine the influence of the sample size and the structure of the population on the power of detecting LD.

Molecular Diversity of the Populations

We observed a higher total molecular allelic richness (8.02 alleles per locus) and average molecular allelic richness of the populations adapted to different MEs (5.7 alleles per locus) than reported in previous SSR studies of maize germplasm, although the Local Southern sizing method used to estimate the fragment sizes represents a conservative allele-calling procedure. Labate et al. (2003) found an average of 6.5 alleles per locus analyzing 461 plants representing a diverse array of U.S. germplasm. Matsuoka et al. (2002) found, on average, 6.9 alleles per locus for 101 maize inbred lines representing three major germplasm sources (Tropical, U.S., and Canadian/European inbreds). The total gene diversity across all populations (0.62) in our study was the same as reported by Matsuoka et al. (2002). The high molecular allelic richness (Table 1) and gene diversity values in our study confirm the broad genetic base of the populations expected from the pedigree data (Table 1).

The allelic richness and number of unique alleles were significantly higher for the tropical populations (6.07 alleles per locus, 86 alleles, respectively) than for populations adapted to the three other MEs (5.86, 5.43, and 5.34 alleles per locus, 37, 23, and 22 alleles). However, the results of the resampled tropical populations obtained from the same number of individuals per population as sampled in each of the subtropical and temperate populations clearly demonstrated the importance of the number of individuals investigated: the more individuals that are sampled, the higher is the probability of detecting rare alleles. The low difference of the pools compared with the populations with respect to average gene diversity (H_S), number of alleles, and number of unique alleles (Table 2) was surprising because (i) the pools have been assessed with a larger effective population size than were the populations and (ii) new material has been regularly introgressed into the pools. Our results indicate that a loss of rare alleles in the populations caused by drift seems to be uncommon and suggests that maintaining back-up pools is not necessary.

Genetic Structure of the Populations

PCoA based on MRD of the populations (Fig. 2) clearly supported the ME structure. PC1, which explained 23.6% of the total variance, revealed a major split between the (i) tropical, (ii) subtropical intermediate-maturity, and (iii) subtropical early-maturity and temperate ME. The position of the four temperate pools between the two groups of the subtropical early-maturity populations (P46, Pl27 vs. P48, Pl30) can be explained by the germplasm base of these populations and pools (Table 1) and the similar selection pressure applied while adapting them to winter maize areas in the subtropics and tropics.

Most of the variation was found within the populations and just a minor part (9% on average) between the populations. The higher G_{ST} values for the tropical, subtropical intermediate-maturity, and subtropical early-maturity populations (0.10, 0.09, and 0.09, respectively) than for the temperate populations (0.07) can be explained by the pedigree information (Table 1). In the temperate populations, many germplasm sources were combined to establish broad-based pools comprising different racial complexes. An analysis of G_{ST} values for individual loci revealed that the following SSRs were associated with the structuring of the germplasm: phi014, phi031, phi053, and phi112. Such a tendency may indicate that the chromosomal regions harboring these SSRs are not selectively neutral. Several studies reported QTL for the anthesis-silking interval in the vicinity of phi014 and phi031 (Ribaut et al., 1996; Veldboom et al., 1994). QTL for days to pollen were reported in chromosomal regions near phi053 and phi112 (CIMMYT, unpublished data). This seems to be an interesting starting point for further fine-scale and association mapping approaches of the underlying genes.

The clustering observed in the tropical populations is largely consistent with the pedigree information (Table 1). Pl24 was formed of Tuxpeño germplasm. P21 was established from seven Tuxpeño races and some families from Pl24. Although P43 was derived from Tuxpeño germplasm, it did not cluster closely with P21, consistent with field data that show high levels of heterosis between P21 and P43. In addition to Tuxpeño germplasm, P22 and P29 contain other materials such as ETO or Cuban flint. However, the results of the PCoA suggest that both populations (P29 and P22) contain mainly Tuxpeño germplasm. P21 and P32 were widely separated in the PCoA, consistent with numerous reports showing substantial heterosis between Tuxpeño and ETO germplasm (Wellhausen, 1978).

In the subtropical intermediate-maturity populations, individuals of P34 and P42 clustered together, consistent with the pedigree information. P42 and P34 both contain ETO germplasm. The latter includes also Cuban flints and Tuxpeño germplasm. The wide distribution of Pl31 over the first and second PCs can be explained by the broad range of germplasm used in its formation (Table 1). Individuals of Pl31 overlapped with individuals of P47 and Pl34, which is again consistent with pedigree information. P47 was formed using 276 half-sibs of Pl32, which itself was established with germplasm from the same sources as Pl31. Individuals of P45 are adjacent to individuals of P33 and have an intersection with them. P45 contains mainly Tuxpeño and U.S. dents but also Cuban flint, the latter being related to Cateto flint from P33 (Goodman and Brown, 1988).

In the subtropical early-maturity populations, two clearly separated clusters were observed: individuals of Pl27 and P46 vs. Pl30 and P48. Individuals of Pl28 were positioned midway between these two groups, which was again in accordance with pedigree information. Pl27

and P46 were both established using flint germplasm from different countries. P48 was generated from 54 half-sib families of Pl30, which was established from dent germplasm from Europe, China, Lebanon, South America and the U.S. Corn Belt. In contrast, Pl28 was developed by mixing flint and dent germplasm from Pl27 and Pl30.

In the temperate populations, the individuals of the four pools were widely spread over the first and second PC and only Pl42 was separated from the three other pools. This reflects nicely the selection history and the establishment of the germplasm (Table 1). Pl42 was formed to introduce tropical germplasm into temperate areas, whereas Pl39, Pl40, and Pl41 were designed to introgress temperate germplasm for the winter maize areas in the subtropics and tropics.

The analysis of the 23 maize populations clearly revealed that most of the genetic diversity is within the populations and just a minor part between the populations. This can be explained by the establishment of the populations and pools, which mostly disregarded racial complexes, and suggests that the applied procedures to handle the broad range of available germplasm was suboptimal with regard to (i) maintaining maximum genetic diversity within the populations and (ii) conserving genetic diversity between the populations. It is rather likely that desired alleles, which occurred with high frequency in just one racial complex, can be lost by mixing different germplasm sources.

Germplasm based on different racial complexes might be useful for the improvement of open–pollinated varieties. However, this germplasm is less suitable for hybrid breeding, where clearly distinct heterotic groups are advantageous (Melchinger, 1999). The reduced genetic diversity among the populations caused by admixture can only be recovered by long-term isolation or reciprocal recurrent selection programs. Therefore, only the few populations based on one racial complex (P21, P32, P33, P42, P43, and Pl24) seem to be suitable for hybrid breeding programs. If no populations based on one racial complex are available for a certain ME, breeders can use either populations not adapted to the ME or landraces in their search for germplasm suitable for hybrid breeding.

ACKNOWLEDGMENTS

The molecular marker analyses of this research were supported by funds from the German "Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung" Project No. 98.7860.4-001-01. The authors thank three anonymous reviewers for their valuable suggestions.

REFERENCES

Chakraborty, R. 1980. Gene-diversity analysis in nested subdivided populations. Genetics 96:721–726.

CIMMYT. 1998. A complete listing of maize germplasm from CIMMYT. Maize Program Special Report. Mexico DF, Mexico.

Dubreuil, P., and A. Charcosset. 1998. Genetic diversity within and among maize populations: A comparison between isozyme and nuclear RFLP loci. Theor. Appl. Genet. 96:577–587.

Elder, J.K., and E.M. Southern. 1987. Computer-aided analysis of one-dimensional restriction fragments gels. p. 165–172 *In* M.J. Bishop and C.J. Rawling (ed.) Nucleid acid and protein sequence analysis—A practical approach. IRL Press, Oxford, UK.

Excoffier, L., and M. Slatkin. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol. Biol. Evol. 12:921–927.

Falconer, D.S., and T.F.C. Mackay. 1996. Introduction to quantitative genetics. 4th ed. Longman Group Ltd., London.

Frisch, M., M. Bohn, and A.E. Melchinger. 2000. Plabsim: Software for simulation of marker-assisted backcrossing. J. Hered. 91:86–87.

Gauthier, P., B. Gouesnard, L. Dallard, R. Redaelli, C. Rebourg, A. Charcosset, and A. Boyat. 2002. RFLP diversity and relationships among traditional European maize populations. Theor. Appl. Genet. 105:91–99.

Goodman, M.M., and C.W. Stuber. 1983. Races of maize: VI. Isozyme variation among races of maize in Bolivia. Maydica 28:169–187.

Goodman, M.M., and W.L. Brown. 1988. Races of Corn. p. 39–74. *In* G.F. Sprague and J.W. Dudley (ed.) Corn and corn improvement. 3rd ed. Agron. Monogr. 18. ASA, CSSA, and SSSA, Madison, WI.

Gower, J.C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325–338.

Guo, S., and E. Thompson. 1992. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48:361–372.

Heckenberger, M., A.E. Melchinger, J.S. Ziegle, L.K. Joe, J.D. Hauser, M. Hutton, and M. Bohn. 2002. Variation of DNA fingerprints among accessions within maize inbred lines with regard to the identification of essentially derived varieties. I. Genetic and technical sources of variation in SSR data. Mol. Breed. 10:181–191.

Ihaka, R., and R. Gentleman. 1996. A language for data analysis and graphics. J. of Computational and Graphical Statistics, Vol. 5. 3:299–314.

Kahler, A.L., A.R. Hallauer, and C.O. Gardner. 1986. Allozyme polymorphisms within and among open-pollinated and adapted exotic populations of maize. Theor. Appl. Genet. 72:592–601.

Kresovich, S., J.G.K. Williams, J.R. Mc Ferson, E.J. Routman, and B.A. Schaal. 1992. Characterization of genetic identities and relationships of *Brassica oleracea* L. via a random amplified polymorphic DNA assay. Theor. Appl. Genet. 85:190–196.

Labate, J.A., K.R. Lamkey, S.H. Mitchell, S. Kresovich, H. Sullivan, and J.S.C. Smith. 2003. Molecular and historical aspects of Corn Belt dent diversity. Crop Sci. 43:80–91.

Labate, J.A., K.R. Lamkey, M. Lee, and W. Woodman. 2000. Hardy-Weinberg and linkage equilibrium estimates in the BSSS and BSCB1 random mated populations. Maydica 45:243–255.

Lynch, M., and B. Walsh. 1997. Genetics and analysis of quantitative traits. p. 413. Sinauer Assoc., Sunderland, MA.

Matsuoka, Y., S.E. Mitchell, S. Kresovich, M. Goodman, and J. Doebley. 2002. Microsatellites in *Zea*– variability, patterns of mutations, and use for evolutionary studies. Theor. Appl. Genet. 104:436–450.

Melchinger, A.E. 1999. Genetic diversity and heterosis. Chapter 10. *In* J.G. Coors and S. Pandey (ed.) The genetics and exploitation of heterosis in crops. CSSA, Madison, WI.

Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

Powell, W., M. Morgante, C. Andre, M. Hanafey, J. Vogel, S. Tingey, and A. Rafalski. 1996. The comparison of RFLP, RAPD, AFLP, and SSR (microsatellite) markers for germplasm analysis. Mol. Breed. 2:225–238.

Remington, D.L., J.M. Thornsberry, Y. Matsuoka, L.M. Wilson, S.R. Whitt, J. Doebley, S. Kresovich, M.M. Goodman, and E.S. Buckler. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. (USA) 98:11479–11484.

Ribaut, J.M., D.A. Hoisington, J.A. Deutsch, C. Jiang, and D. Gonzalez de Leon. 1996. Identification of quantitative trait loci under drought conditions in tropical maize. 1. Flowering parameters and the anthesis-silking interval. Theor. Appl. Genet. 92:905–914.

Ron Parra, J., and A.R. Hallauer. 1997. Utilisation of exotic maize germplasm. Plant Breed. Rev. 14:165–187.

Saghai-Maroof, M.A., K.M. Soliman, R. Jorgenson, and R.W. Allard. 1984. Ribosomal DNA spacer length polymorphisms in barley: Mendelian inheritance, chromosomal location and population dynamics. Proc. Natl. Acad. Sci. (USA) 81:8014–8018.

Schneider, S., D. Roessli, and L. Excoffier. 2000. Arlequin, ver. 2.0:

A software of population genetics data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Slatkin, M., and L. Excoffier. 1996. Testing for linkage disequilibrium in genotypic data using the EM algorithm. Heredity 76:377–383.

Smith, J.S.C., E.C.L. Chin, H. Shu, O.S. Smith, S.J. Wall, M.L. Senior, S.E. Mitchell, S. Kresovich, and J. Ziegle. 1997. An evaluation of utility of SSR loci as molecular markers in maize (*Zea mays* L.): Comparisons with data from RFLPs and pedigree. Theor. Appl. Genet. 95:163–173.

Snedecor, G.W., and W.G. Cochran. 1980. Statistical methods. Iowa State Univ. Press, Ames.

Stuber, C.W., R.H. Moll, M.M. Goodman, H.E. Schaffer, and B.E. Weir. 1980. Allozyme frequency changes associated with selection for increased grain yield in maize (*Zea mays* L.). Genetics 93:225–236.

Vasal, S.K., H.S. Cordova, S. Pandey, and G. Srinivasan. 1999. Tropical maize and heterosis. Chapter 34. *In* J.G. Coors and S. Pandey (ed.) The genetics and exploitation of heterosis in crops. CSSA, Madison, WI.

Veldboom, L.R., M. Lee, and W.L. Woodman. 1994. Molecular marker-facilitated studies in an elite maize population: 1. Linkage analysis and determination of QTL for morphological traits. Theor. Appl. Genet. 88:7–16.

Warburton, M.L., X. Xianchun, J. Crossa, J. Franco, A.E. Melchinger, M. Frisch, M. Bohn, and D. Hoisington. 2002. Genetic characterization of CIMMYT inbred maize lines and open pollinated populations using large scale fingerprinting methods. Crop Sci. 42:1832–1840.

Wellhausen, E.J. 1978. Recent developments in maize breeding in the tropics. p. 59–91. *In* D.B. Walden (ed.) Maize breeding and genetics. John Wiley & Sons, New York.

Wright, S. 1978. Evolution and genetics of populations, Vol. IV. p. 91. The Univ. of Chicago Press, Chicago.

Genetic Distance Based on Simple Sequence Repeats and Heterosis in Tropical Maize Populations

J. C. Reif, A. E. Melchinger,* X. C. Xia, M. L. Warburton, D. A. Hoisington, S. K. Vasal, G. Srinivasan, M. Bohn, and M. Frisch

ABSTRACT

Heterotic groups and patterns are of fundamental importance in hybrid breeding of maize (*Zea mays* L.). The major goal of this study was to investigate the relationship between heterosis and genetic distance determined with simple sequence repeat (SSR) markers. The objectives of our research were to (i) compare the genetic diversity within and between seven tropical maize populations, (ii) test alternative hypotheses on the relationship between panmictic midparent heterosis (PMPH) and genetic distances determined with SSR markers, and (iii) evaluate the use of SSR markers for grouping of germplasm and establishing heterotic patterns in hybrid breeding of tropical maize. Published data of a diallel of seven tropical maize populations evaluated for agronomic traits in seven environments were reanalyzed to calculate PMPH in population hybrids. In addition, 48 individuals from each population were sampled and assayed with 85 SSR markers covering the entire maize genome. A total of 532 alleles in the 7 × 48 genotypes assayed were detected. The analysis of molecular variance (AMOVA) revealed that 89.8% of the variation was found within populations and only 10.2% between populations. The correlation between PMPH and the squared modified Roger's distance (MRD) based on SSR markers was significantly positive ($P < 0.05$) only for grain yield ($r = 0.63$). With SSR analyses, it was possible to assign Population 29 (Pop29) to the established Heterotic Group A and propose new heterotic groups (Pop25, Pop43). We conclude that SSR markers provide a powerful tool for grouping of germplasm and are a valuable complementation to field trials for identifying groups with satisfactory heterotic response.

GENETIC DIVERSITY in maize plays a key role for future breeding progress. The development of molecular markers provides a tool for assessing the genetic diversity at the DNA level in plant species (Melchinger and Gumber, 1998). In particular, SSR markers show potential for large-scale DNA fingerprinting of maize genotypes due to the high level of polymorphism detected (Smith et al., 1997), their analyses by automated systems (Sharon et al., 1997), and their high accuracy and repeatability (Heckenberger et al., 2002).

Most evidence in maize suggests that the genetic basis of heterosis is partial to complete dominance (Hallauer et al., 1988; Stuber et al., 1992). Overdominance has long been discussed as the basis of heterosis (East, 1936; Crow, 1948). However, many data supporting overdominance presumably resulted from pseudooverdominance, arising from dominant alleles in repulsion phase linkage (Stuber et al., 1992; Crow, 1999). Epistasis, particularly between linked loci, may also be an explanation for heterosis in maize (Cockerham and Zeng, 1996). No data exclude the possibility of all three mechanisms contributing to heterosis, albeit in different proportions.

Lamkey and Edwards (1999) coined the term *panmictic midparent heterosis* to describe the deviation in performance between a population cross and the mean of its two parent populations in Hardy-Weinberg equilibrium. Quantitative genetic theory shows that in the absence of epistasis and two alleles per locus, PMPH is a function of the product of the dominance effect and the square of the difference in gene frequencies at the respective locus (Falconer and Mackay, 1996, p. 255), which corresponds to the square of the MRD (Melchinger, 1999). In fact, a linear increase in PMPH with increasing genetic distance (Hypothesis 1) was hypothesized in a diallel of U.S. maize populations (Moll et al., 1962).

In contrast, experimental data reported by Moll et al. (1965) in a study with tropical maize populations of diverse geographic origin suggest that PMPH increases with increasing genetic distance only up to an optimum level but thereafter decreases in extremely wide crosses (Hypothesis 2). The authors explained this by fertility distortion in wide crosses and epistatic interactions of genes. While Moll et al. (1962, 1965) inferred the genetic distance from the geographic origin of the populations, to our knowledge no attempts have been made to verify or falsify the above hypotheses with more reliable data based on molecular markers.

The choice of heterotic groups is fundamental in hybrid breeding of maize (Melchinger and Gumber, 1998). While heterotic patterns in temperate maize have been established more than 50 yr ago, a clearly defined heterotic pattern does not exist in the tropical maize of the CIMMYT germplasm. Therefore, before embarking on a hybrid breeding program, CIMMYT conducted several diallel studies for identifying populations showing not only good per se performance but also high heterosis in their crosses (Beck et al., 1990; Crossa et al., 1990; Vasal et al., 1992a,b,c). Genetic distances based on molecular markers have been suggested as a tool for grouping of similar germplasm as a first step in identifying promising heterotic patterns (Melchinger, 1999).

The major goal of this study was to investigate the relationship between heterosis and genetic distance de-

J.C. Reif, A.E. Melchinger, M. Bohn, and M. Frisch, Inst. of Plant Breeding, Seed Sci., and Population Genetics, Univ. of Hohenheim, 70593 Stuttgart, Germany; X.C. Xia, M.L. Warburton, D.A. Hoisington, S.K. Vasal, and G. Srinivasan, Int. Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641 06600 Mexico D.F., Mexico. Received 7 June 2002. *Corresponding author (melchinger@uni-hohenheim.de).

Published in Crop Sci. 43:1275–1282 (2003).

Abbreviations: ANOVA, analysis of variance; AMOVA, analysis of molecular variance; CIMMYT, International Maize and Wheat Improvement Center; GCA, general combining ability; MRD, modified Roger's distance; PC, principal coordinate; PCoA, principal coordinate analysis; PMPH, panmictic midparent heterosis; Pop, population; SCA, specific combining ability; SSR, simple sequence repeat.

termined with SSR markers. The objectives of our research were to (i) compare the genetic diversity within and between seven tropical maize populations, (ii) test alternative hypotheses on the relationship between PMPH and genetic distances determined with SSR markers, and (iii) evaluate the use of SSR markers for grouping of germplasm and establishing heterotic patterns for hybrid breeding of tropical maize.

MATERIALS AND METHODS

Field Trials

The field experiments were previously described in detail by Vasal et al. (1992a). Briefly, their investigation involved six tropical late white maize populations and one gene pool developed by CIMMYT (Table 1). The seven maize populations were crossed in a 7 × 7 diallel mating design at Poza Rica, Mexico, in the 1985 winter season. All possible 21 crosses were made in both directions using bulked pollen of each parent population. Seeds from each cross and its reciprocal were bulked to represent a particular cross. Seed increase of each parent population was done simultaneously by random mating to ensure Hardy-Weinberg equilibrium.

The parents and their crosses were evaluated in field trials for grain yield, days to silking, and plant height at seven locations (Tlaltizapán, Poza Rica, Silao, Tlacomulco, and Obregón in Mexico; Palmira in Colombia; and Nakornsawan in Thailand) during 1985-1986. The experimental design was a randomized complete block design with three replications at each location. The experimental unit consisted of two 5-m rows spaced 75 cm and a plant density of ≈53 333 plants ha^{-1}. All rows were hand-harvested and grain yield was calculated from dry ear weight at harvest assuming 80% shelling and adjusted to 155 g kg^{-1} grain moisture.

Simple Sequence Repeat Analyses

From each of the seven populations, 48 randomly chosen individuals were analyzed separately. The seeds used for extracting DNA were from the same selection cycle as the populations tested in the field trials; however, the populations were multiplied repeatedly by CIMMYT's maize genebank since 1985.

DNA was extracted employing the CTAB procedure (Clarke et al., 1989). The 85 SSR markers were chosen from the MaizeDB database (*http://nucleus. agron.missouri.edu/cgi-bin/ssr_bin.pl*) based on repeat unit and bin location to provide uniform coverage of the entire maize genome. The SSRs were multiplexed for maximum efficiency. Fragments were separated using acrylamide gels run on an ABI 377 automatic DNA sequencer. Fragment sizes were calculated with GeneScan 3.1 (Perkin Elmer/Applied Biosystems) using the Local Southern sizing method; allele identity was assigned using Genotyper 2.1 (Perkin Elmer/Applied Biosystems) and the two inbred lines CML51 and CML292 as control. Data have been stored in the MaizeDB database (*http://nucleus. agron.missouri.edu/cgi-bin/ssr_bin.pl*).

Statistical Analyses

Analyses of variance (ANOVA) were computed for the three plant traits. A mixed linear model was used with the assumption that effects of entries were fixed and all other effects were considered random. Following Analysis III of Gardner and Eberhart (1966), the sums of squares and degrees of freedom (27 *df*) for entries were orthogonally partioned into the contrast between parents *vs.* crosses (1 *df*), the variation among populations (6 *df*), and the variation among crosses (20 *df*) with a further subdivision into general combining ability (GCA) and specific combining ability (SCA) effects. A corresponding subdivision was made on the entry × environment interaction sums of squares. Entry mean squares were tested by *F* tests for significance by using the corresponding entry × environment mean squares. Entry × environment mean squares were tested for significance by using the pooled error mean square. The PMPH of each cross was calculated as the difference between the F_1 mean and the respective midparent mean across all environments.

The gene diversity (D) based on SSR data was calculated for each population according to Weir (1996, p. 151):

$$D = 1 - \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{a_i} p_{ij}^2, \qquad [1]$$

where p_{ij} is the frequency of the jth allele at the ith marker, a_i is the number of alleles at the ith marker, and m refers to the number of markers. In addition to D, we used the AMOVA to divide the genetic variation into components attributable to the variance between and within populations (Michalakis and Excoffier, 1996).

We calculated the MRD between two populations or individuals (Wright, 1978, p. 91; Goodman and Stuber, 1983) as:

Table 1. Description of the seven CIMMYT tropical late maize populations used in this study.

Population or pool	Name; selection cycle; Heterotic Group	Germplasm description
Pool24	**Tropical Late White Dent; C_{21}; A**	**Mainly based on Tuxpeño germplasm but includes also some materials from Central America, the Carribean, and Zaire. White dent grain type. Tolerant to ear and stalk rots. Selected for resistance to fall armyworm.**
Pop21	**Tuxpeño-1; C_5; A**	**Composed of seven Tuxpeño races plus some familes from Pool 24. White dent grain type. Excellent standability and relatively short plant type. Fairly tolerant to most foliar diseases.**
Pop22	**Mezcla Tropical Blanc; C_6; A**	**Broad genetic base, including Tuxpeño, ETO Blanco, Antigua, and Central American germplasm. White dent-semident grain type. Improved for downy mildew resistance in Thailand and the Philippines.**
Pop25	**Blanco-Cristalino-3; C_0; B**	**Derived from tropical late white flint Pool 23. Composed of white flint selections from crosses among materials from Mexico, Colombia, the Caribbean, Central America, India, Thailand, and the Philippines. White flint grain type. Improved for husk cover and resistance to ear and stalk rot as well as root and stalk lodging.**
Pop29	**Tuxpeño Caribe; C_5; unassigned**	**Broad genetic base including Tuxpeño, Cuban flints, and ETO. White dent grain type. Improved for reduced plant height, stalk and root lodging, and husk cover.**
Pop32	**ETO Blanco; C_5; B**	**Developed in Colombia with germplasm from South America, Cuba, Mexico, and the U.S. cornbelt. White flint grain type. Improved for shorter plant type at CIMMYT.**
Pop43	**La Posta; C_5; unassigned**	**Tuxpeño synthetic composed of 16 S_1 lines. White grain type. Improved for resistance to streak virus in Nigeria.**

Table 2. Means (above diagonal) and panmictic midparent heterosis (below diagnoal) for grain yield, days to silking, and plant height of seven CIMMYT tropical late white maize populations and their crosses averaged across data from seven environments during 1985 and 1986.

	Populations						
	Pool24	**Pop21**	**Pop22**	**Pop25**	**Pop29**	**Pop32**	**Pop43**
	Grain yield, Mg ha^{-1}						
per se	**6.36**	**6.66**	**7.12**	**6.31**	**6.51**	**5.96**	**7.05**
Pool24		**7.22**	**6.90**	**6.80**	**6.78**	**6.56**	**6.98**
Pop21	**0.71**		**7.34**	**7.40**	**6.98**	**7.15**	**7.83**
Pop22	**0.16**	**0.45**		**6.92**	**7.21**	**7.55**	**7.55**
Pop25	**0.47**	**0.92**	**0.21**		**6.78**	**6.68**	**7.07**
Pop29	**0.37**	**0.40**	**0.40**	**0.37**		**7.34**	**7.06**
Pop32	**0.64**	**0.84**	**1.01**	**0.55**	**1.11**		**7.40**
Pop43	**0.10**	**0.98**	**0.47**	**0.39**	**0.28**	**0.90**	***0.49***†
	Plant height, cm						
per se	**217.0**	**217.9**	**212.9**	**205.9**	**204.1**	**217.1**	**234.9**
Pool24		**224.9**	**217.2**	**216.6**	**213.9**	**218.5**	**227.1**
Pop21	**7.5**		**219.6**	**216.7**	**210.8**	**223.8**	**226.8**
Pop22	**2.3**	**4.2**		**208.3**	**212.1**	**216.3**	**216.7**
Pop25	**5.2**	**4.8**	**−1.1**		**215.1**	**213.0**	**220.8**
Pop29	**6.1**	**−0.2**	**3.6**	**10.1**		**211.2**	**216.9**
Pop32	**−0.5**	**6.3**	**1.3**	**1.5**	**0.6**		**229.6**
Pop43	**−9.4**	**0.4**	**−7.2**	**0.4**	**−2.6**	**3.6**	***8.8***†
	Days to silking, d						
per se	**68.0**	**69.8**	**67.4**	**66.4**	**68.3**	**69.1**	**70.4**
Pool24		**69.4**	**67.3**	**67.9**	**68.0**	**67.8**	**69.8**
Pop21	**0.5**		**68.2**	**67.7**	**69.4**	**68.8**	**68.9**
Pop22	**−0.4**	**−0.4**		**66.3**	**66.4**	**67.1**	**67.1**
Pop25	**0.7**	**−0.4**	**−0.6**		**66.5**	**66.9**	**67.9**
Pop29	**−0.3**	**0.4**	**−1.5**	**−0.9**		**68.0**	**69.2**
Pop32	**−0.7**	**−0.7**	**−1.2**	**−0.9**	**−0.7**		**68.9**
Pop43	**−1.3**	**−1.2**	**−1.8**	**−0.5**	**−0.2**	**−0.9**	***1.4***†

† LSD0.05 of the means.

$$\text{MRD} = \sqrt{\frac{1}{2m}\sum_{i=1}^{m}\sum_{j=1}^{a_i}(p_{ij} - q_{ij})^2}, \qquad [2]$$

where p_{ij} and q_{ij} are allele frequencies of the jth allele at the ith marker in the two entries under consideration and a_i and m as defined above. Standard errors of MRD estimates were obtained by using a bootstrap procedure with resampling over markers and individuals.

Associations among the populations were revealed with principal coordinate analysis (PCoA) (Gower, 1966) based on MRD estimates. Multiple regression analysis was used to study the relationship between PMPH and squared modified Roger's distance (MRD^2). The PCoA was performed with the statistical software R (Ihaka and Gentleman, 1996) and multiple regression analysis with the statistical software SAS (SAS Institute, 1988).

RESULTS

Agronomic Trials

The combined ANOVA showed highly significant ($P < 0.01$) differences among the 28 entries (7 populations, 21 crosses) for all three traits, but no significant genotype × environment interactions (Table 2 of Vasal et al., 1992a). The comparison of parents vs. crosses, which provides a measure for average PMPH, was significant ($P < 0.01$) only for grain yield and amounted to 0.56 Mg ha^{-1}. Grain yield differed significantly ($P < 0.01$) among the seven parent populations as well as among the 21 crosses and ranged from 5.96 Mg ha^{-1} (Pop32) to 7.12 Mg ha^{-1} (Pop22) for the parent populations and from 6.56 Mg ha^{-1} (Pop32 × Pool 24) to 7.83 Mg ha^{-1} (Pop21 × Pop43) for the crosses (Table 2). The variation among the crosses was mainly due to significant ($P < 0.01$) GCA effects, whereas SCA effects were not significant for any trait.

Maximum PMPH for grain yield was observed in cross Pop29 × Pop32 with 1.11 Mg ha^{-1}, although it was not the top yielding cross. Minimum PMPH was observed in cross Pop43 × Pool24 with 0.10 Mg ha^{-1}.

Simple Sequence Marker Data

The 85 SSR primers generated a total of 532 alleles in the 336 genotypes (7 populations × 48 individuals) analyzed. The number of alleles per marker across all seven populations was on average 6.3 and ranged from 2 to 16 (Table 3). Gene diversity D within the seven populations ranged from 0.503 to 0.580 with a mean of 0.539 (Table 3). Values of MRD between pairs of populations averaged 0.258 and ranged from 0.203

Table 3. Gene diversity D within populations, average number ($\bar{a}$) and standard deviation σ_a of alleles per population.

	Population							
Statistic	**Pool24**	**Pop21**	**Pop22**	**Pop25**	**Pop29**	**Pop32**	**Pop43**	**Total**
D	**0.559**	**0.548**	**0.535**	**0.527**	**0.580**	**0.518**	**0.503**	**0.593**
$\bar{a}$	**4.247**	**4.259**	**4.226**	**4.000**	**4.294**	**3.541**	**3.553**	**6.259**
σ_a	**2.029**	**1.814**	**1.679**	**1.766**	**1.792**	**1.593**	**1.687**	**2.583**

Table 4. Modified Roger's distances between populations (above diagonal) and their standard error (below diagonal).

	Population						
Population	**Pool24**	**Pop21**	**Pop22**	**Pop25**	**Pop29**	**Pop32**	**Pop43**
Pool24		0.219	0.203	0.224	0.216	0.270	0.248
Pop21	0.016		0.222	0.272	0.236	0.305	0.286
Pop22	0.014	0.016		0.250	0.233	0.284	0.268
Pop25	0.021	0.024	0.021		0.259	0.263	0.278
Pop29	0.017	0.015	0.016	0.023		0.285	0.274
Pop32	0.017	0.019	0.020	0.023	0.023		0.318
Pop43	0.019	0.018	0.018	0.026	0.019	0.021	

(Pool24 × Pop22) to 0.318 (Pop32 × Pop43) with significant ($P < 0.01$) differences between MRD estimates (Table 4). The AMOVA revealed that 89.8% of the molecular genetic variance was found within populations and 10.2% between populations (Table 5).

In the PCoA based on MRD estimates for the populations, the first three principal coordinates (PC) explained 27.3, 22.1, and 15.8% of the total variation, respectively (Fig. 1). Pop21, Pop22, Pop29, and Pool24 were clearly separated from Pop32 and Pop25 with respect to the first principal coordinate (PC1). Pop43 and Pop25 were separated from the other populations with respect to PC2 and PC3. Principal coordinate analysis based on individual plants also resulted in a clear separation between a cluster consisting of Pop21, Pop22, Pop29, and Pool24 and a cluster comprising Pop25, Pop32, and Pop43 (Fig. 2).

Relationship between Panmictic Midparent Heterosis and Marker Data

The MRD^2 was plotted against PMPH of grain yield, plant height, and days to silking (Fig. 3) and analyzed with multiple regression. The MRD^2 was significantly correlated with PMPH for grain yield ($r = 0.63$; $P < 0.01$) and negatively for days to silking ($r = -0.44$; $P < 0.05$) and plant height ($r = -0.13$). Neither the quadratic nor the cubic regression model gave a significantly better fit to the data than the linear regression (data not shown).

DISCUSSION

CIMMYT's maize germplasm bank contains about 8000 accessions of tropical maize for use in breeding. Breeding efforts at CIMMYT in the early 1960s and 1970s were focused on population improvement via recurrent selection and, therefore, emphasized formation of genetically broad-based populations and pools disregarding heterotic patterns and combining ability (Vasal et al., 1999). Their mixed genetic constitution makes the task of assigning them to genetically diverse and complementary heterotic groups difficult. To achieve this goal, germplasm originally developed by intermating genetically diverse races were grouped according to ecology, grain color, and maturity. The groups were tested in diallel designs, each involving six to 10 populations or pools. On the basis of their performance data, the populations were categorized (Vasal et al., 1999). Pop21, Pop22, and Pool24 were assigned to Heterotic Group A, while Pop25 and Pop32 were allotted to Heterotic Group B. Pop29 and Pop43 have not yet been assigned to these or other heterotic groups.

Table 5. Analysis of molecular variance of the seven tropical maize populations analyzed with 85 SSR markers.

Source of variation	df	SS	Variance components	% variation
Among populations	6	1 443.8	2.3	10.2
Within populations	665	13 430.6	20.2	89.8
Total	671	14 874.4	22.5	100.0

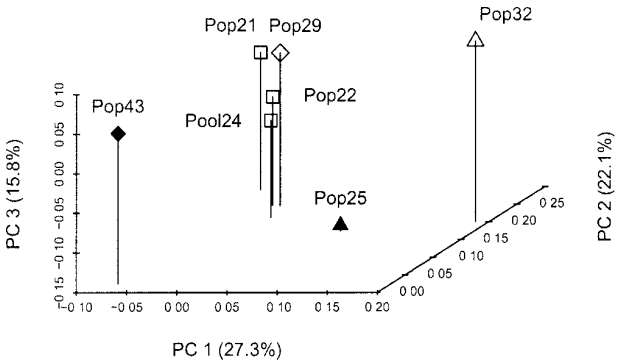

Fig. 1. Principal coordinate analysis of the seven tropical maize populations based on modified Roger's distance. PC1, PC2, and PC3 are the first, second, and third principal coordinate, respectively. Heterotic Group A (Pop21, Pop22, and Pool24), Heterotic Group B (Pop25, Pop32), and populations not yet assigned to heterotic groups (Pop29, Pop43) are shown.

Genetic Diversity among and within the Populations

In this study, we found on average across the seven populations 6.3 alleles per marker. Lu and Bernardo

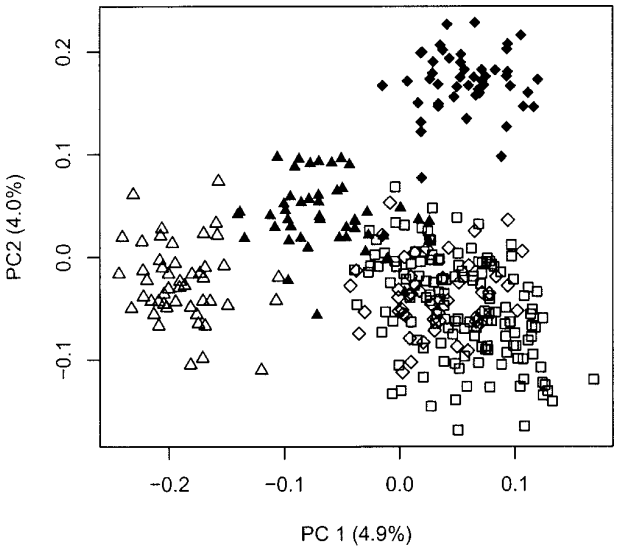

Fig. 2. Principal coordinate analysis of individuals from seven tropical maize populations based on modified Roger's distance. PC1 and PC2 are the first and second principal coordinate, respectively. Heterotic Group A (Pop21, Pop22, and Pool24, open squares), Heterotic Group B (Pop25, filled triangles; Pop32, open triangles), and populations not yet assigned to heterotic groups (Pop29, open diamonds; Pop43, filled diamonds) are shown.

(2001) detected for 40 U.S. inbred lines an average of 4.9 alleles using 83 SSR markers. Senior et al. (1998) reported an average of five alleles in their study with 94 elite maize inbreds, representative of the diversity in the U.S. maize germplasm, and 70 SSR markers. Hence, the total number of alleles per marker was higher in our study than previously reported in the literature. This and the high average number of alleles per population (Table 3) in our study suggests a broad genetic base of the seven populations.

Pop29 had the highest gene diversity *D* followed by Pool24 and Pop21 (Table 3). This is consistent with pedigree information (Table 1) because the populations have been established using a wide range of germplasm. The lowest *D* value observed for Pop43 is also in accordance with its pedigree, because it was generated from 16 S_1 lines including only Tuxpeño germplasm. Ranking of the populations based on *D* was almost identical with their ranking based on the average number of alleles per marker (rank correlation $r_s = 0.93$; $P < 0.01$). Altogether, the high percentage (89.8%) of the molecular variance revealed by the AMOVA (Table 5) within populations is in harmony with the broad genetic base of the materials used for their synthesis (Table 1). Since related germplasm such as various sources from Tuxpeño or ETO entered different populations, it was also not surprising to find only a minor variance between populations (Table 5). A more detailed analysis of the population subdivision with test statistics of the AMOVA was not possible, because this would require knowledge of the gametic phase for linked loci (Michalakis and Excoffier, 1996), which cannot be determined from SSR analyses of heterozygous individuals.

Correlation between MRD^2 and Panmictic Midparent Heterosis

We investigated the correlation between PMPH and MRD^2 because quantitative genetic theory suggests a linear relationship between both measures under certain assumptions (Falconer and Mackay, 1996, p. 255). This is in harmony with related studies on midparent heterosis in crosses of inbred lines (see Melchinger et al., 1991; Boppenmaier et al., 1993), where the commonly employed Roger's distance (1972) is equal to MRD^2 (Melchinger, 1993). A low correlation between PMPH and MRD^2 can be attributable to several causes: (i) a poor association between heterozygosity estimated from marker data and heterozygosity at quantitative trait loci affecting the trait examined, (ii) a poor association between heterozygosity and heterosis at quantitative trait loci in the crosses examined (Charcosset et al., 1991), (iii) existence of multiple alleles (Cress, 1966), and (iv) epistasis (Moll et al., 1965).

The low correlations between MRD^2 and PMPH for plant height and days to silking were mostly due to small PMPH estimates for these traits (Table 2). By comparison, the corresponding correlation for grain yield was surprisingly high ($r = 0.63$; $P < 0.01$). This is consistent with the relative large contribution of SCA effects to the total sums of squares, which accounted for 33% of the genetic variation among crosses for this trait (Vasal et al., 1992a). In accordance with quantitative genetic theory (Melchinger, 1999) the correlation of MRD^2 was lower with hybrid performance ($r = 0.41$; $P < 0.05$) than with PMPH for grain yield ($r = 0.63$; $P < 0.01$). On the basis of a literature survey with single crosses produced from inbreds, Melchinger (1999) pointed out that only intragroup crosses show a correlation between parental genetic distance and midparent heterosis, but not intergroup crosses. However, a closer examination of the graph between MRD^2 and PMPH (Fig. 3) did not provide any clue in this direction.

While Hypothesis 1 postulates a linear relation between MRD^2 and PMPH, under Hypothesis 2 a quadratic or cubic regression is expected to fit the data better than linear regression. However, in our study neither a quadratic nor a cubic regression model gave a significantly better fit to the data than linear regression. This is in accordance with the graphs shown in Fig. 3. Consequently, our results confirm Hypothesis 1 for the tropical maize germplasm investigated here.

A decrease in PMPH of genetically very distant populations is generally attributed to the lack of coadaption between both allelic and nonallelic combinations of genes from the two parental haploid genomes, resulting in reduced or negative dominance and negative epistatic effects, respectively (Falconer and Mackay, 1996, p. 255). A major reason for the absence of an optimum in the relationship between genetic distance and PMPH in our study could be that all populations (Table 1) were more or less well adapted to the test environments. In addition, we did not include extremely wide crosses, as was the case in the experiment of Moll et al. (1965).

For hybrid breeding, Melchinger and Gumber (1998) recommended the following criteria for the choice of heterotic patterns: (i) high mean performance and large genetic variance in the hybrid population; (ii) high per se performance and good adaption of the parent populations to the target region(s); (iii) low inbreeding depression, if hybrids are produced from inbreds. Under Hypothesis 1 (PMPH increases with increasing genetic distance), genetic distance could be used as a further criterion for the identification of heterotic patterns. Considering all four criteria, the following promising heterotic patterns can be suggested: (i) Heterotic Group A with Heterotic Group B; (ii) Pop43 with Heterotic Group A or B; (iii) Pop29 with Heterotic Group B or Pop43.

Grouping of Germplasm

We chose the MRD as genetic distance measure because of its mathematical and genetic properties. In particular, it is an Euclidean distance, which is an often-overlooked prerequisite for most multivariate analysis methods (Jacquard, 1974, p. 465). Furthermore, in the absence of epistasis and two alleles per locus, PMPH is a linear function of the product of the dominance effect and the square of the MRD (Melchinger, 1999).

Principal coordinate analysis based on MRD revealed very clearly a major split between the populations from

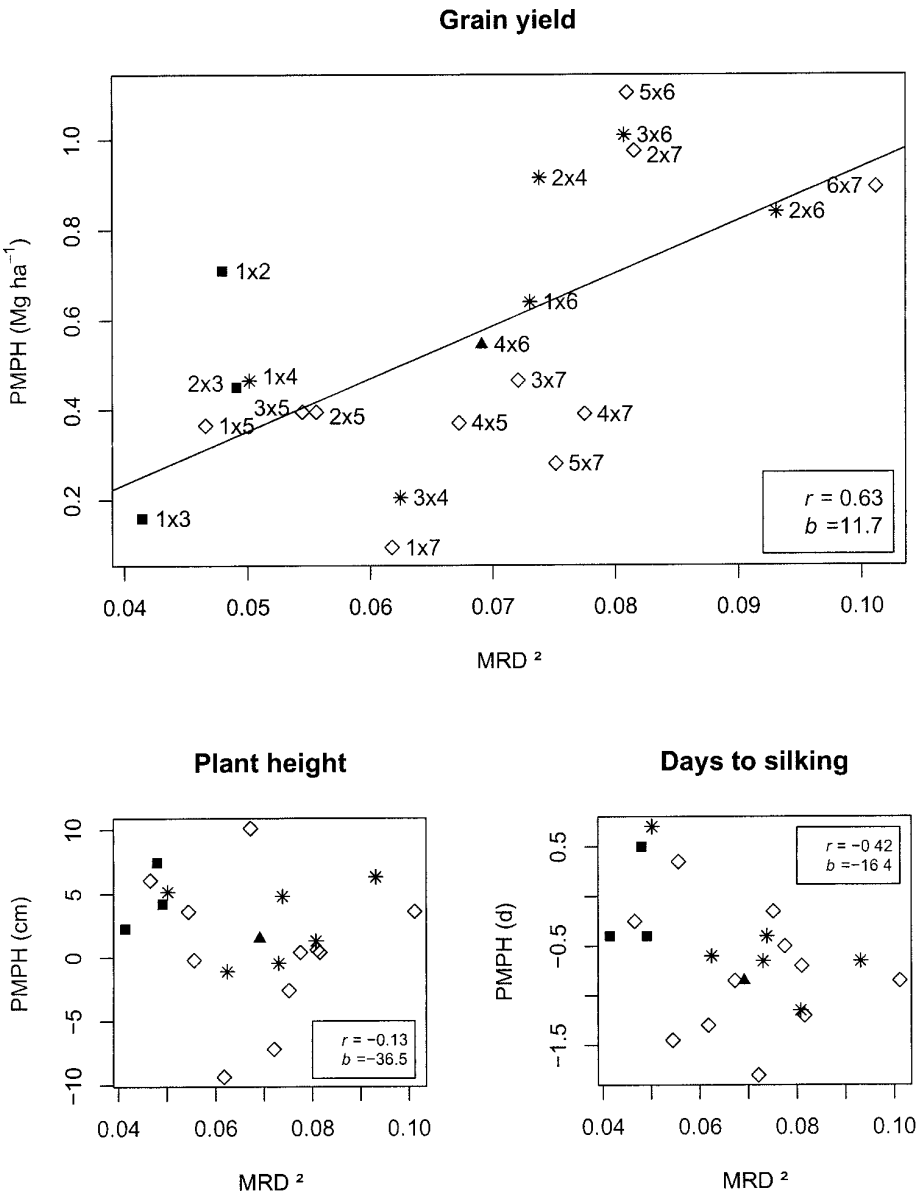

Fig. 3. Relationship between squared Roger's distance (MRD^2) and panmictic midparent heterosis (PMPH) for grain yield, plant height, and days to silking. Intrapool crosses within Heterotic Group A (filled squares) and Group B (filled triangles), interpool crosses between A and B (*), and miscellaneous (open diamonds) are shown. 1 = Pool24, 2 = Pop21, 3 = Pop22, 4 = Pop25, 5 = Pop29, 6 = Pop32, 7 = Pop43; *r* is the correlation coefficient and *b* the slope coefficient.

Heterotic Group A and Pop32 (Fig. 1). Pop25 is separated from the other populations by PC3 and had an average MRD at the population level to Heterotic Group A of 0.24 and to Pop32 of 0.26. The assignment of Pop25 to Heterotic Group B together with Pop32 originally based on testcross data was not supported by our molecular data. This could be interpreted as an indicator that Pop25 should have been established as a separate Heterotic Group C. The values of PMPH (Table 2) support this hypothesis in that Pop25 had a low average PMPH with Heterotic Group A. In addition, PCoA accurately portrayed the relationship of Pop43 to Heterotic Group A and B. It is closer to Heterotic Group A ($\overline{MRD}$ = 0.26) than to Heterotic Group B ($\overline{MRD}$ = 0.29), but the distance from Pop43 to Heterotic Group A was higher than the average distance between Heterotic Groups A and B. This together with the diallel analysis suggests classification of Pop43 as a separate Heterotic Group D. According to the PCoAs (Fig. 1, 2), Pop29 could be assigned to Heterotic Group A, because it had a smaller average MRD to Heterotic Group A (0.22) than to B (0.26). The diallel analysis supports this suggestion.

In conclusion, classification of the seven populations based on SSR data mostly confirmed the results from the diallel data except the assignment of Pop25 to Heterotic

Group B. Furthermore, it was possible to assign Pop29 to the established Heterotic Groups A and to propose new heterotic groups (Pop25, Pop43). When a large number of germplasm exists but no established heterotic groups are available, genetically similar germplasm can be identified with molecular markers. On basis of this information, field trials can be planned more efficiently. Thus, by using molecular data to focus the search for heterotic groups on a smaller number of promising heterotic patterns and evaluating these intensively, breeders should arrive at a more economic and solid approach for making this important decision at the beginning of a hybrid breeding program.

CONCLUSIONS

The results of this study suggest that molecular marker-based analyses, and in particular SSR technology, offers a reliable and effective means of assessing genetic diversity within and between maize populations. The AMOVA revealed a high within population variance, as expected from the origin and genetic background of these populations. For the establishment of heterotic groups to be used in hybrid breeding, a higher variance between populations would have been advantageous because this should result in higher PMPH and, consequently, a higher performance of crosses between them.

Simple sequence repeat markers provide a valuable tool for grouping of germplasm and are a good complementation to field trials for identifying groups of genetically similar germplasm. Consequently, field trials for identification of promising heterotic patterns can be planned more efficiently based on prior information obtained from SSR analyses.

ACKNOWLEDGMENTS

The molecular marker analysis of this research were supported by funds from the German "Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung" Project No. 98.7860.4-001-01. J.C. Reif was supported by the "Vater und Sohn Eiselen Stiftung", Ulm, Germany, for carrying out parts of this research at CIMMYT in Mexico. The authors thank three anonymous reviewers for their valuable suggestions.

REFERENCES

Beck, D.L., S.K. Vasal, and J. Crossa. 1990. Heterosis and combining ability of CIMMYT's tropical early and intermediate maturity maize (*Zea mays* L.) germplasm. Maydica 35:279–285.

Boppenmaier, J., A.E. Melchinger, G. Seitz, H.H. Geiger, and R.G. Herrmann. 1993. Genetic diversity for RFLPs in European maize inbreds. III. Performance of crosses within versus between heterotic groups for grain traits. Plant Breed. 111:217–226.

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

Clarke, B.C., L.B. Moran, and R. Appels. 1989. DNA analyses in wheat breeding. Genome 32:334–339.

Cockerham, C.C., and Z.B. Zeng. 1996. Design III with marker loci. Genetics 143:1437–1456.

Cress, C.E. 1966. Heterosis of the hybrid related to gene frequency differences between two populations. Genetics 53:269–274.

Crossa, J., S.K. Vasal, and D.L. Beck. 1990. Combining ability estimates of CIMMYT tropical late yellow maize germplasm. Maydica 35:273–278.

Crow, J.F. 1948. Alternative hypotheses of hybrid vigor. Genetics 33:477–487.

Crow, J.F. 1999. Dominance and overdominance. p. 49–58. *In* J.G. Coors and S. Pandey (ed.) The genetics and exploitation of heterosis in crops. ASA, CSSA, and SSSA, Madison, WI.

East, E.M. 1936. Heterosis. Genetics 21:375–397.

Falconer, D.S., and T.F. Mackay. 1996. Introduction to quantitative genetics. 4th ed. Longman Group, London.

Gardner, C.O., and S.A. Eberhart. 1966. Analysis and interpretation of the variety cross diallel and related populations. Biometrics 22: 439–452.

Goodman, M.M., and C.W. Stuber. 1983. Races of maize: VI. Isozyme variation among races of maize in Bolivia. Maydica 28:169–187.

Gower, J.C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325–388.

Hallauer, A.R., W.A. Russell, and K.R. Lamkey. 1988. Corn breeding. p. 463–564. *In* G.F. Sprague and J.W. Dudley (ed.) Corn and corn improvement. 3rd ed. Agron. Monogr. 18. ASA, CSSA, and SSSA, Madison, WI.

Heckenberger, M., A.E. Melchinger, J.S. Ziegle, L.K. Joe, J.D. Hauser, M. Hutton, and M. Bohn. 2002. Variation of DNA fingerprints among accessions within maize inbred lines with regard to the identification of essentially derived varieties. I. Genetic and technical sources of variation in SSR data. Mol. Breed. 10:181–191.

Ihaka, R., and R. Gentleman. 1996. A language for data analysis and graphics. J. Comput. Graph. Stat. 5:299–314.

Jacquard, A. 1974. The genetic structure of populations. p. 465–466. *In* Biomathematics. Vol. 5. Springer Verlag, Berlin.

Lamkey, K.R., and J.W. Edwards. 1999. Quantitative genetics of heterosis. p. 31–48. *In* J.G. Coors and S. Pandey (ed.) The genetics and exploitation of heterosis in crops. ASA, CSSA, and SSSA, Madison, WI.

Lu, H., and R. Bernardo. 2001. Molecular marker diversity among current and historical maize inbreds. Theor. Appl. Genet. 103: 613–617.

Melchinger, A.E. 1993. Use of RFLP markers for analysis of genetic relationship among breeding materials and prediction of hybrid performance. p. 621–628. *In* D.R. Buxton et al. (ed.) International crop science I. CSSA, Madison, WI.

Melchinger, A.E. 1999. Genetic diversity and heterosis. p. 99–118. *In* J.G. Coors and S. Pandey (ed.) The genetics and exploitation of heterosis in crops. ASA, CSSA, and SSSA, Madison, WI.

Melchinger, A.E., and R.K. Gumber. 1998. Overview of heterosis and heterotic groups in agronomic crops. p. 29–44. *In* K.R. Lamkey and J.E. Staub (ed.) Concepts and breeding of heterosis in crop plants. CSSA Spec. Publ. 25. CSSA, Madison, WI.

Melchinger, A.E., M.M. Messmer, M. Lee, W.L. Woodman, and K.R. Lamkey. 1991. Diversity and relationships among U.S. maize inbreds revealed by restriction fragment length polymorphisms. Crop Sci. 31:669–678.

Michalakis, Y., and L. Excoffier. 1996. A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. Genetics 142:1061–1064.

Moll, R.H., J.H. Longquist, J.V. Fortuna, and E.C. Johnson. 1965. The relation of heterosis and genetic divergence in maize. Genetics 52:139–144.

Moll, R.H., W.S. Salhuana, and H.F. Robinson. 1962. Heterosis and genetic diversity in variety crosses of maize. Crop Sci. 2:197–198.

SAS Institute. 1988. SAS language guide for personal computers. Release 6.03 ed. SAS Inst., Cary, NC.

Senior, M.L., J.P. Murphy, M.M. Goodman, and C.W. Stuber. 1998. Utility of SSRs for determining genetic similarities and relationships in maize using an agarose gel system. Crop Sci. 38:1088–1098.

Sharon, E.M., S. Kresovich, C.A. Jester, C.J. Hernandez, and A.K. Szewc-McFadden. 1997. Application of multiplex PCR and fluorescence-based, semi-automated allele sizing technology for genotyping plant genetic resources. Crop Sci. 37:617–624.

Smith, J.S.C., E.C.L. Chin, H. Shu, O.S. Smith, S.J. Wall, M.L. Senior, S.E. Mitchell, S. Kresovich, and J. Ziegle. 1997. An evaluation of utility of SSR loci as molecular markers in maize (*Zea mays* L.):

comparisons with data from RFLPs and pedigree. Theor. Appl. Genet. 95:163–173.

Stuber, C.W., S.E. Lincoln, D.W. Wolff, T. Helentjaris, and E.S. Lander. 1992. Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. Genetics 132:823–839.

Vasal, S.K., H. Cordova, S. Pandey, and G. Srinivasan. 1999. Tropical maize and heterosis. p. 363–373. *In* J.G. Coors and S. Pandey (ed.) The genetics and exploitation of heterosis in crops. ASA, CSSA, and SSSA, Madison, WI.

Vasal, S.K., G. Srinivasan, D.L. Beck, J. Crossa, S. Pandey, and C. de Leon. 1992a. Heterosis and combining ability of CIMMYT's tropical late white maize germplasm. Maydica 37:217–223.

Vasal, S.K., G. Srinivasan, J. Crossa, and D.L. Beck. 1992b. Heterosis and combining ability of CIMMYT's subtropical and temperate early maturity maize germplasm. Crop Sci. 32:884–890.

Vasal, S.K., G. Srinivasan, S. Pandey, H.S. Cordova, G.C. Han, and F. Gonzalez. 1992c. Heterotic patterns of ninety two white tropical CIMMYT maize lines. Maydica 37:259–270.

Weir, B.S. 1996. Genetic data analysis II. 2nd ed. Sinauer Associates, Sunderland, MA.

Wright, S. 1978. Evolution and genetics of populations. Vol. IV. Univ. of Chicago Press, Chicago, IL.

Theor Appl Genet (2003) 107:947–957
DOI 10.1007/s00122-003-1333-x

J. C. Reif · A. E. Melchinger · X. C. Xia · M. L. Warburton · D. A. Hoisington · S. K. Vasal · D. Beck · M. Bohn · M. Frisch

Use of SSRs for establishing heterotic groups in subtropical maize

Received: 18 November 2002 / Accepted: 19 April 2003 / Published online: 27 June 2003

Abstract Heterotic groups and patterns are of fundamental importance in hybrid breeding. The objectives of our research were to: (1) investigate the relationship of simple sequence repeats (SSR) based genetic distances between populations and panmictic midparent heterosis (PMPH) in a broad range of CIMMYT maize germplasm, (2) evaluate the usefulness of SSR markers for defining heterotic groups and patterns in subtropical germplasm, and (3) examine applications of SSR markers for broadening heterotic groups by systematic introgression of other germplasm. Published data of two diallels and one factorial evaluated for grain yield were re-analyzed to calculate the PMPH in population hybrids. Additionally, 20 pools and populations widely used in CIMMYT's breeding program were assayed with 83 SSR markers covering the entire maize genome. Correlations of squared modified Roger's distance (MRD^2) and PMPH were mostly positive and significant, but adaption problems caused deviations in some cases. For intermediate- and early-maturity subtropical germplasm, two heterotic groups could be suggested consisting of a flint and dent composite. We concluded that the relationships between the populations obtained by SSR analyses are in excellent agreement with pedigree information. SSR markers are a valuable complementation to field trials for identifying heterotic groups and can be used to introgress exotic germplasm systematically.

Keywords Heterotic groups · SSRs · Heterosis · Mega-environment · Genetic distance

Communicated by F. Salamini

J. C. Reif · A. E. Melchinger (✉) · M. Frisch
Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany
e-mail: melchinger@uni-hohenheim.de
Fax: +49-711-4592343

M. L. Warburton · D. A. Hoisington · S. K. Vasal · D. Beck
CIMMYT, International Wheat and Maize Improvement Center. Apdo. Postal 6-641, 06600 Mexico DF, Mexico

X. C. Xia
National Maize Improvement Center of China, China Agricultural University, 100094, Beijing, China

M. Bohn
Crop Science Department, Universiy of Illinois, 1102 South Goodwin Avenue, Urbana, IL 61801, USA

Introduction

Recognition of heterotic patterns among genetically divergent groups of germplasm is fundamental in hybrid breeding for maximum exploitation of heterosis (Hallauer et al. 1988). Lamkey and Edwards (1999) coined the term panmictic midparent heterosis (PMPH) for the difference between a hybrid population and the mean of its two parent populations in Hardy-Weinberg equilibrium. Under the assumptions of two alleles per locus and no epistasis, PMPH is a function of the dominance effect at each locus and the square of the difference in allele frequency between the populations (Falconer and Mackay 1996); the latter corresponds to the square of the modified Roger's distance (MRD^2).

Using the geographic origin as a crude indicator for the genetic distance, Moll et al. (1962) in their study with U.S. maize observed a linear increase in PMPH with increasing genetic distances. In contrast, experimental data reported by Moll et al. (1965) in a study with tropical and U.S. maize populations suggested an increase of PMPH with increasing genetic distance only up to an optimum level, but a decrease in extremely wide crosses. The authors explained this decline by fertility distortion in wide crosses, adaptation problems and epistatic interactions of genes. The relationship between mid-parent heterosis of single-cross hybrids and the genetic distance of their parental inbreds, determined with molecular markers, were investigated both in theory (Charcosset and Essioux 1994) and numerous experiments with maize and other crops (Brummer 1999). Melchinger (1999) pointed out that only intragroup crosses show a correlation between parental genetic distance and midparent hetero-

sis, but for intergroup hybrids, heterosis is at best only loosely correlated with the parental genetic distance.

If heterosis of hybrids increases monotonically with increasing genetic distance of the parents, genetic distances based on molecular markers should be a useful tool for establishing promising heterotic groups and patterns (Melchinger and Gumber 1998). Introgression of exotic germplasm is often suggested for increasing the genetic differences between opposite heterotic populations with an expected increase in heterotic response (Beck et al. 1991; Vasal et al. 1992a, b; Ron Parra and Hallauer 1997).

Over the past 35 years, breeders at the International Maize and Wheat Improvement Center (CIMMYT) have developed numerous germplasm pools, populations, and open-pollinated varieties (OPV) based on mixtures of germplasm originating from various backgrounds (CIMMYT 1998). A series of combining ability studies was conducted to determine heterotic relationships among CIMMYT populations and pools. Several of the populations demonstrated good general combining ability, and various promising heterotic patterns were identified (Crossa et al. 1990; Beck et al. 1991; Vasal et al. 1992a, b). However, no conclusions were drawn about clearly defined heterotic groups. With the establishment of a hybrid breeding program, the question of suitable heterotic groups becomes relevant for subtropical maize germplasm (Vasal et al. 1999).

The objectives of our research were to: (1) investigate the relationship of simple sequence repeat (SSR) based genetic distances between populations and PMPH in a broad range of CIMMYT maize germplasm, (2) evaluate the usefulness of SSR markers for defining heterotic groups and patterns in subtropical germplasm, and (3) examine applications of SSR markers for broadening heterotic groups by systematic introgression of other germplasm.

Materials and methods

For reducing the large collection of germplasm from CIMMYT's gene bank, to a size which can be handled efficiently for breeding purposes, more than 100 populations were established using germplasm from different sources. Additionally, 30 broad-based back-up pools were formed to reduce the danger of narrowing down the genetic basis in tropical and subtropical maize (CIMMYT 1998). We investigated using molecular markers 20 of these pools and populations (further referred to as populations) (Table 1), which had previously been included in published field experiments.

Field experiments

Experiment 1 comprised a complete diallel of five subtropical early-maturity and two temperate populations (Pop46, 48, and Pool27, 28, 30, 40, 42) described in detail by Vasal et al. (1992a). Experiment 2 included a complete diallel of seven intermediate-maturity subtropical populations (Pop33, 34, 42, 45, 47 and Pool31, 34) and two temperate adapted populations (Pool39, 41) published by Beck et al. (1991). Experiment 3 comprised factorial crosses (Design-II, Comstock and Robinson 1948) of four intermediate-maturity subtropical populations (Pop42, Pop45, Pop47, Pool34) mated with four tropical populations (Pop22, Pop25, Pop32, Pop43) described in detail by Vasal et al. (1992c). In addition to the hybrid populations, all parent populations were included in each experiment. Experiment 1 was evaluated in five subtropical (four Mexican, one Turkish) and 17 temperate (16 U.S., one Canadian) environments. Experiment 2 was tested in five subtropical environments in Mexico. Experiment 3 was evaluated in six environments in Mexico and Colombia. The experimental design for the three experiments was a randomized complete block design with three replications in each environment. All crosses in both reciprocal forms were produced at Poza Rica, Mexico, in the 1985 winter season using bulked pollen of each parent population. Seeds from each cross and its reciprocal were bulked to represent a particular cross. Seed increase of each parent population was done simultaneously by random mating to ascertain Hardy-Weinberg equilibrium.

The parents and their crosses were evaluated for grain yield. In the subtropical environments, the experimental unit consisted of two 5-m rows spaced 75 cm, and a plant density of approximately 53,333 plants ha^{-1}. In temperate environments, plot size and plant density varied; at most sites, the experimental unit was two rows either 3.05 or 6.10 m in length, spaced either 0.76 or 0.91 m apart. Final stands ranged from 53,333 to 87,700 plants ha^{-1}. For the subtropical environments all rows were hand-harvested and grain yield (mg ha^{-1}) was calculated at 80% of the ear weight adjusted to 155 g kg^{-1} of moisture. For all temperate environments, plots were machine harvested and shelled grain weight was adjusted to 155 g kg^{-1} of moisture.

SSR analyses

Twenty one randomly chosen individuals from each of the 16 subtropical and temperate populations, and 48 individuals from the four tropical populations, were analyzed separately. DNA was extracted from plants grown from seed increases of the original populations tested in the field trials.

DNA was extracted employing a modified CTAB procedure (Saghai-Maroof et al. 1984). The 83 SSR markers used in the study were chosen from the MaizeDB database (http://nucleus.agron.missouri.edu/cgi-bin/ssr_bin.pl) based on the repeat unit and bin location to provide uniform coverage of the entire maize genome. Primers and PCR conditions were described in detail by Warburton et al. (2002). Briefly, SSRs were multiplexed for maximum efficiency. Fragments were separated using acrylamide gels run on an ABI 377 automatic DNA sequencer. Fragment sizes were calculated with GeneScan 3.1 (Perkin Elmer/Applied Biosystems) using the Local Southern sizing method (Elder and Southern 1987); allele identity was assigned using Genotyper 2.1 (Perkin Elmer/Applied Biosystems) and the two inbred lines CML51 and CML292 as a control. Data have been stored in the MaizeDB database (http://nucleus.agron.missouri.edu/cgi-bin/ssr_bin.pl).

Statistical analyses

The three experiments were analyzed separately. Analyses of variance (ANOVA) for grain yield were computed for each mega-environment (ME) separately (Experiment 1: subtropical and temperate MEs; Experiment 2: subtropical ME; and Experiment 3: tropical, subtropical, and transition/mid-altitude MEs). Analyses III of Gardner and Eberhart (1966) were carried out for Experiment 1 and 2 and a Design II analysis (Comstock and Robinson 1948) for Experiment 3.

Entry mean squares were tested for significance by *F*-tests by using the corresponding entry × environment mean squares. Entry × environment mean squares were tested for significance by using the pooled error mean square. PMPH of each cross was calculated as the difference between the F_1 mean and the respective midparent mean for each ME.

Table 1 Description of the 20 CIMMYT maize populations used in this study

Population/Pool	Cycle	Experiment	Germplasm description
Tropical			
Pop22	6	3	Includes Tuxpeño and ETO Blanco germplasm, and germplasm from Central America
Pop25	0	3	Is composed of white flint selections from crosses among germplasm from Mexico, Columbia, the Caribbean, Central America, India, Thailand and the Philippines
Pop32	5	3	Is based on white flint germplasm from South America, Cuba, Mexico and the U.S.Cornbelt
Pop43	5	3	Is a Tuxpeño synthetic composed of 16 S_1 lines
Subtropical intermediate-maturity			
Pop33	2	2	Contains mainly Argentinian (Cateto) flints
Pop34	5	2	Includes Cuban flints, ETO, Tuxpeño, and germplasm from the U.S. Cornbelt, India and Nepal
Pop42	4	2 and 3	Is an advanced generation of ETO selected for short-plant type and crossed with Illinois Cornbelt components
Pop45	3	2 and 3	Includes U.S. Cornbelt germplasm, Tuxpeño, Cuban flints, Puerto Rico composite, and collections from the Dominican Republic
Pop47	2	2 and 3	Consists largely of Tuxpeño germplasm plus some U.S. Cornbelt lines
Pool31	14	2	Is a broadbased pool including white flint segregates from Ecuador, Argentina, India, Mexico, Pool32, and Pool33, but contains also germplasm from Mexico, U.S. Cornbelt, Brazil, Uruguay, Argentina, China, Pakistan, Yugoslavia, Lebanon, Guatemala, Venezuela, Peru, Cuba, and the Dominican Republic
Pool34	20	2 and 3	Includes germplasm from the Mexican lowlands and highlands, the U.S. Cornbelt, southern USA, Puerto Rico, Pakistan, Hungary, China, Peru, Pakistan, Lebanon, Nicaragua and Guatemala
Subtropical early-maturity			
Pop46	1	1	Represents a superior flint fraction (240 half-sib families) of Pool 29, which is based on germplasm from Europe, Lebanon, U.S. Cornbelt, China, Indonesia and South America.
Pop48	5	1	Is composed of dents from U.S. Cornbelt germplasm, southern European germplasm and 54 half-sib families from Pool 30
Pool27	20	1	Includes flint germplasm from the USA, China, Lebanon, Pakistan and several European countries
Pool28	14	1	Is based on crosses between white dent segregates from Pool27 and Hungarian germplasm from Pool 30, and various other germplasms
Pool30	15	1	Made up of dent germplasm from Europe, China, Lebanon, Mexico, South America and the U.S. Cornbelt
Temperate			
Pool39	12	2	Contains germplasm from the tropical lowlands and highlands, subtropical and temperate areas
Pool40	12	1	Is based on germplasm from Europe
Pool41	12	2	Includes predominantly U.S. Cornbelt germplasm plus germplasm from China, Korea and Lebanon
Pool42	12	1	Is based on germplasm from Mexico, Peru, Bolivia,Pakistan, Hungary, the USA and Yemen

We calculated the modified Roger's distance (MRD) between two populations (Wright 1978, pp 91; Goodman and Stuber 1983) as:

$$\mathrm{MRD} = \sqrt{\frac{1}{2m}\sum_{i=1}^{m}\sum_{j=1}^{a_i}(p_{ij} - q_{ij})^2}. \quad (1)$$

Here, p_{ij} and q_{ij} are the allele frequencies of the jth allele at the ith marker in the two populations under consideration, a_i is the number of alleles at the ith marker, and m refers to the number of markers. Standard errors of MRD estimates were obtained by using a bootstrap procedure with re-sampling over markers and individuals within populations. Following Melchinger et al. (1990), the squared modified Roger's distance (MRD^2) was partitioned into general ($GMRD^2$) and specific squared modified Roger's distances ($SMRD^2$) analogous to the subdivision of agronomic data into GCA and SCA effects. Pearson correlation coefficients (r) were calculated for MRD^2 and $SMRD^2$ with F_1 performance, PMPH and SCA effects. Significance tests of r were performed by using tabulated values based on Fisher (1921) z transformation. The polymorphic-index content (PIC) for each SSR marker was determined as described by Smith et al. (1997).

A principal coordinate analysis (PCoA, Gower 1966) was calculated separately for each experiment based on the matrix of MRD values. Heterotic groups were defined by using the k-means clustering algorithm (Hartigan and Wong 1979), which assigns populations to k clusters such that the within-cluster sum of squares is minimized. The predefined number k of clusters was choosen based on: (1) pedigree information, (2) information from breeders, and (3) the results from PCoA. All analyses were carried out with the

950

Table 2 Means (above diagonal) and panmictic midparent heterosis (PMPH, below diagonal) for grain yield in different mega-environments (ME) and modified Roger's distance (MRD) between populations (above diagonal) and their standard error (SE, below diagonal) of seven CIMMYT's maize populations and their crosses evaluated in Experiment 1

Pop.	Pop46	Pop48	Pool27	Pool28	Pool30	Pool40	Pool42
	Subtropical ME Mg ha^{-1}						
per se	4.50	4.69	4.88	4.99	4.41	3.73	3.26
Pop46		4.89	4.82	4.95	5.17	4.33	4.32
Pop48	0.29		5.42	5.26	4.93	4.45	4.40
Pool27	0.13	0.64		4.92	5.18	4.25	4.37
Pool28	0.21	0.42	–0.02		5.15	4.18	4.39
Pool30	0.72	0.38	0.54	0.45		4.46	4.38
Pool40	0.22	0.24	–0.05	–0.18	0.39		3.80
Pool42	0.44	0.43	0.30	0.27	0.55	0.31	*0.42*[a]
	Temperate ME Mg ha^{-1}						
per se	3.70	4.93	3.80	4.30	4.95	3.80	3.45
Pop46		4.98	4.06	4.28	4.88	3.88	3.90
Pop48	0.67		4.82	5.03	4.96	4.56	4.38
Pool27	0.31	0.46		4.45	4.70	4.23	4.10
Pool28	0.28	0.42	0.40		4.71	4.27	4.17
Pool30	0.56	0.02	0.33	0.09		4.70	4.24
Pool40	0.13	0.19	0.43	0.22	0.33		3.66
Pool42	0.33	0.19	0.48	0.30	0.04	0.04	*0.34*[a]
	MRD (above diagonal) and SE (below diagonal)						
Pop46		0.294	0.234	0.226	0.248	0.247	0.239
Pop48	0.022		0.301	0.257	0.224	0.247	0.256
Pool27	0.024	0.024		0.220	0.269	0.253	0.239
Pool28	0.021	0.023	0.024		0.217	0.214	0.220
Pool30	0.021	0.022	0.022	0.022		0.213	0.232
Pool40	0.021	0.020	0.023	0.021	0.020		0.222
Pool42	0.021	0.023	0.021	0.024	0.022	0.022	

[a] LSD (0.05) of the means

Plabsim software (Frisch et al. 2000), which is implemented as an extension to the statistical software R (Ihaka and Gentleman 1996).

An analysis of molecular variance (AMOVA) (Michalakis and Excoffier 1996) was performed to divide the molecular genetic variance into components attributable to the variance between and within populations using the software package Arlequin (Schneider et al. 2000).

Results

For all three experiments highly significant ($P < 0.01$) differences among the entries, parents, crosses, and parents vs crosses were observed in all MEs (Beck et al. 1991; Vasal et al. 1992a, c). GCA effects were highly significant ($P < 0.01$) in all cases except the transition/mid-altitude MEs in Experiment 3. SCA effects were significant ($P < 0.05$) in Experiment 1 for the temperate ME and highly significant ($P < 0.01$) in Experiment 3 for the subtropical ME.

Experiment 1

Average grain yield in the subtropical ME ranged for the parent populations from 3.26 Mg ha^{-1} (Pool42) to 4.99 Mg ha^{-1} (Pool28) and for the crosses from 3.80 Mg ha^{-1} (Pool40 × Pool42) to 5.42 Mg ha^{-1} (Pop48 × Pool27) (Table 2). PMPH for grain yield was maximum in Pop46 × Pool30 (0.72 Mg ha^{-1}) and minimum in Pool28 × Pool40 (–0.18 Mg ha^{-1}). In the temperate ME, average grain yields for parents and their crosses were 4.13 Mg ha^{-1} and 4.43 Mg ha^{-1}, respectively. Pool30 (4.95 Mg ha^{-1}) and Pop48 (4.93 Mg ha^{-1}) had the highest grain yields among the parents. Average grain yields for the crosses ranged from 3.66 Mg ha^{-1} (Pool40 × Pool42) to 5.03 Mg ha^{-1} (Pop48 × Pool28). PMPH ranged from 0.02 Mg ha^{-1} (Pop48 × Pool30) to 0.67 Mg ha^{-1} (Pop46 × Pop48) and averaged 0.29 Mg ha^{-1}.

Experiment 2

Average grain yield in the subtropical ME ranged from 4.61 Mg ha^{-1} (Pool41) to 7.21 Mg ha^{-1} (Pop42) for the parents and from 4.91 Mg ha^{-1} (Pool39 × Pool41) to 7.87 Mg ha^{-1} (Pop42 × Pop47) for the crosses (Table 3). PMPH averaged 0.38 Mg ha^{-1} with a maximum of 0.92 Mg ha^{-1} (Pop33 × Pop45) and a minimum of –0.09 Mg ha^{-1} (Pop45 × Pool39).

Experiment 3

Average grain yield for the crosses ranged from 5.62 Mg ha^{-1} (Pop32 × Pool34) to 7.43 Mg ha^{-1} (Pop43 × Pop42) for tropical ME, from 6.11 Mg ha^{-1} (Pop25 × Pool34) to 8.03 Mg ha^{-1} (Pop22 × Pop42, Pop43 × Pop42) for subtropical ME, and from 6.23 Mg ha^{-1} (Pop32 × Pool34) to 7.97 Mg ha^{-1} (Pop22 × Pop47) for the transition/mid-altitude ME (Table 4). PMPH averaged 0.59, 0.78 and 0.53 Mg ha^{-1} for the tropical, subtropical and transition/mid-altitude MEs, respectively.

951

Table 3 Means (above diagonal) and panmictic midparent heterosis (PMPH, below diagonal) for grain yield in the temperate mega-environment and modified Roger's distance (MRD) between populations (above diagonal) and their standard error (SE, below diagonal) of nine CIMMYT's maize populations and their crosses evaluated in Experiment 2

Pop.	Pop33	Pop34	Pop42	Pop45	Pop47	Pool31	Pool34	Pool39	Pool41
	Mg ha^{-1}								
per se	5.77	6.60	7.21	6.36	7.01	6.11	6.19	5.16	4.61
Pop33		6.77	6.89	6.98	6.96	6.12	6.17	5.64	5.64
Pop34	0.59		7.40	7.13	7.16	6.64	7.13	6.34	6.08
Pop42	0.40	0.50		7.47	7.87	7.03	7.13	6.38	6.57
Pop45	0.92	0.65	0.69		7.03	6.32	6.62	5.67	5.42
Pop47	0.57	0.36	0.76	0.35		7.06	6.87	6.31	6.47
Pool31	0.18	0.28	0.37	0.09	0.50		6.37	5.94	5.86
Pool34	0.19	0.74	0.43	0.35	0.27	0.22		5.79	5.62
Pool39	0.18	0.46	0.19	–0.09	0.23	0.31	0.11		4.91
Pool41	0.45	0.48	0.66	–0.07	0.66	0.50	0.22	0.03	*0.67*[a]
	MRD (above diagonal) and SE (below diagonal)								
Pop33		0.244	0.264	0.237	0.257	0.251	0.242	0.228	0.256
Pop34	0.024		0.236	0.292	0.268	0.272	0.281	0.277	0.305
Pop42	0.021	0.021		0.284	0.281	0.278	0.270	0.272	0.278
Pop45	0.021	0.023	0.022		0.273	0.245	0.223	0.229	0.230
Pop47	0.022	0.023	0.021	0.021		0.261	0.261	0.276	0.289
Pool31	0.026	0.027	0.025	0.027	0.027		0.269	0.264	0.278
Pool34	0.024	0.024	0.021	0.021	0.022	0.026		0.230	0.260
Pool39	0.021	0.021	0.020	0.019	0.021	0.024	0.022		0.212
Pool41	0.023	0.021	0.020	0.020	0.021	0.024	0.023	0.020	

[a] LSD (0.05) of the means

Table 4 Means and panmictic midparent heterosis (PMPH) for grain yield in different mega-environments (ME) and modified Roger's distances (MRD) between populations and their standard error (SE) of tropical × subtropical crosses and parents evaluated in Experiment 3

Pop.	Pool34	Pop42	Pop45	Pop47	*per se*	Pool34	Pop42	Pop45	Pop47
	Grain yield (Mg ha^{-1})					PMPH (Mg ha^{-1})			
Tropical ME									
Pop22	6.14	6.65	6.26	6.06	6.65	0.90	0.61	0.45	0.11
Pop25	5.64	6.79	6.28	6.45	6.27	0.59	0.94	0.66	0.69
Pop32	5.62	5.66	6.45	6.06	6.13	0.64	–0.13	0.90	0.37
Pop43	6.34	7.43	6.93	6.35	7.25	0.80	1.09	0.82	0.10
per se	3.84	5.44	4.97	5.25	*0.74*[a]				
Subtropical ME									
Pop22	7.38	8.03	6.99	7.27	7.56	1.18	0.68	0.38	0.07
Pop25	6.11	7.70	6.63	7.31	6.77	0.30	0.75	0.41	0.51
Pop32	6.87	7.09	7.17	7.12	6.00	1.45	0.52	1.34	0.71
Pop43	7.16	8.03	7.43	7.27	6.73	1.37	1.10	1.23	0.49
per se	4.85	7.14	5.67	6.83	*0.66*[a]				
Transition/mid-altitude ME									
Pop22	6.50	6.45	6.65	7.97	6.83	0.42	–0.41	0.69	1.34
Pop25	6.38	6.51	6.64	7.01	6.46	0.49	–0.17	0.86	0.57
Pop32	6.23	7.21	6.83	7.06	6.73	0.20	0.40	0.92	0.48
Pop43	6.68	7.64	6.55	7.06	6.71	0.66	0.84	0.65	0.49
per se	5.33	6.89	5.10	6.43	*0.92*[a]				
MRD/SE									
	MRD					SE			
Pop22	0.274	0.308	0.299	0.298		0.022	0.021	0.021	0.021
Pop25	0.290	0.300	0.294	0.284		0.024	0.023	0.024	0.023
Pop32	0.295	0.278	0.321	0.307		0.024	0.022	0.022	0.019
Pop43	0.324	0.326	0.327	0.316		0.025	0.021	0.024	0.023

[a] LSD (0.05) of the means

SSR marker data

The 83 SSR primers generated a total of 641 alleles in the 528 genotypes analyzed. The number of alleles per marker across the 20 populations was on average 7.7 and ranged from 2 to 17. PIC values for the SSR loci ranged from 0.10 to 0.85, with an average of 0.60. MRD between pairs of populations for Experiment 1, 2 and 3 averaged 0.241, 0.260, 0.303, and ranged from 0.213 (Pool30 × Pool40) to 0.301 (Pop48 × Pool27), 0.212 (Pool39 × Pool41) to 0.305 (Pop34 × Pool41), and 0.274 (Pop22 × Pool34) to 0.326 (Pop43 × Pop42), respectively (Tables 2, 3 and 4).

952

Table 5 Analysis of molecular variance (AMOVA) of the populations from the three experiments based on 83 SSR markers

Source of variation	*df*	Sum of Squares	Variance components	Percentage of variation
	Experiment 1			
Among populations	6	399.8	1.2	6.3
Within populations	285	5,000.5	17.5	93.7
Total	*291*	*5,400.3*	*18.7*	*100.0*
	Experiment 2			
Among populations	8	581.8	1.3	7.0
Within populations	369	6,474.4	17.5	93.0
Total	*377*	*7,056.2*	*18.8*	*100.0*
	Experiment 3			
Among populations	7	1,302.2	10.4	11.6
Within populations	544	10,274.5	22.3	88.4
Total	*551*	*11,576.7*	*32.8*	*100.0*

Table 6 Correlations of squared modified Roger's distance (MRD^2) and specific squared modified Roger's distance ($SMRD^2$) based on 83 SSR markers obtained for the parent populations in maize with various parameters (Y) from the analyses of generation means of the grain yield data for different mega-environments of three experiments

Parameter Y	Experiment					
	1 ST[a]	1 TR[a]	2 ST[a]	3 TR[a]	3 ST[a]	3 TM[a]
	r(MRD^2, Y)					
F_1 performance	0.43*	0.40	0.47**	0.64**	0.37	0.14
SCA effects[b]	0.34	0.28	0.35*	0.29	0.18	–0.07
PMPH[c]	0.37	0.56**	0.53**	0.33	0.43	0.18
	r($SMRD^2$, Y)					
F_1 performance	0.16	0.16	0.12	0.28	0.17	–0.08
SCA effects[b]	0.55**	0.47*	0.45**	0.51	0.31	–0.12
PMPH[c]	0.32	0.36	0.34*	0.40	0.19	–0.09

*, ** Significant at the 0.05 and 0.01 levels of probability, respectively
[a] TR, ST and TM refers to tropical, subtropical, and transition mid-altitude mega-environments, respectively
[b] Specific combining ability
[c] PMPH is the panmictic midparent heterosis

PCoA was peformed separately for each experiment (Fig. 1). In Experiment 1, principle coordinate (PC) 1 clearly separated: (1) Pool27 and Pop46, from (2) Pool30 and Pop48, whereas Pool28, Pool40 and Pool42 were positioned in between these two groups. In Experiment 2, PC1 separated: (1) Pop34 and Pop42, from (2) Pop33, Pop45, Pool34, Pool39 and Pool41. PC2 separated these two groups from Pop47 and Pool31. The populations investigated in Experiment 3 formed two clearly separated clusters: (1) Pop22, Pop25, Pop32 and Pop43, and (2) Pop42, Pop45, Pop47 and Pool34.

For all three experiments the AMOVA revealed only a small proportion ($\leq$11.6%) of the molecular variance among populations and the major proportion within populations (Table 5).

Correlations of MRD^2 and $SMRD^2$ with F_1 performance, SCA effects and PMPH estimated from the field data, were positive except for Experiment 3 in the transition/midaltitude ME (Table 6). SCA effects were more closely correlated with $SMRD^2$ than MRD^2. In contrast, PMPH was more closely related with MRD^2 than $SMRD^2$ and highly significant ($P < 0.01$) in two instances (Fig. 2).

Discussion

For hybrid breeding, Melchinger and Gumber (1998) recommended the following criteria for the choice of heterotic patterns: (1) high mean performance and large genetic variance in the hybrid population; (2) high per se performance and good adaption of the parent population to the target region(s); and (3) low inbreeding depression, if hybrids are produced from inbred lines. The main focus of this study was to investigate the use of SSR markers for the grouping of germplasm and the identification of promising heterotic patterns before evaluating the germplasm in intensive field trials.

Descriptive statistics

In this study, we found on average across the 20 populations a higher number of alleles per marker (7.7) than reported by Lu and Bernardo (2001) investigating 40 U.S. inbred lines with 83 SSR markers (4.9), and Senior et al. (1998) evaluating 94 elite U.S. maize inbreds with 70 SSR markers (5.0). This can be explained by the broad germplasm base captured in the 20 populations and the diverse origin of their ancestors (Table 1). In contrast to

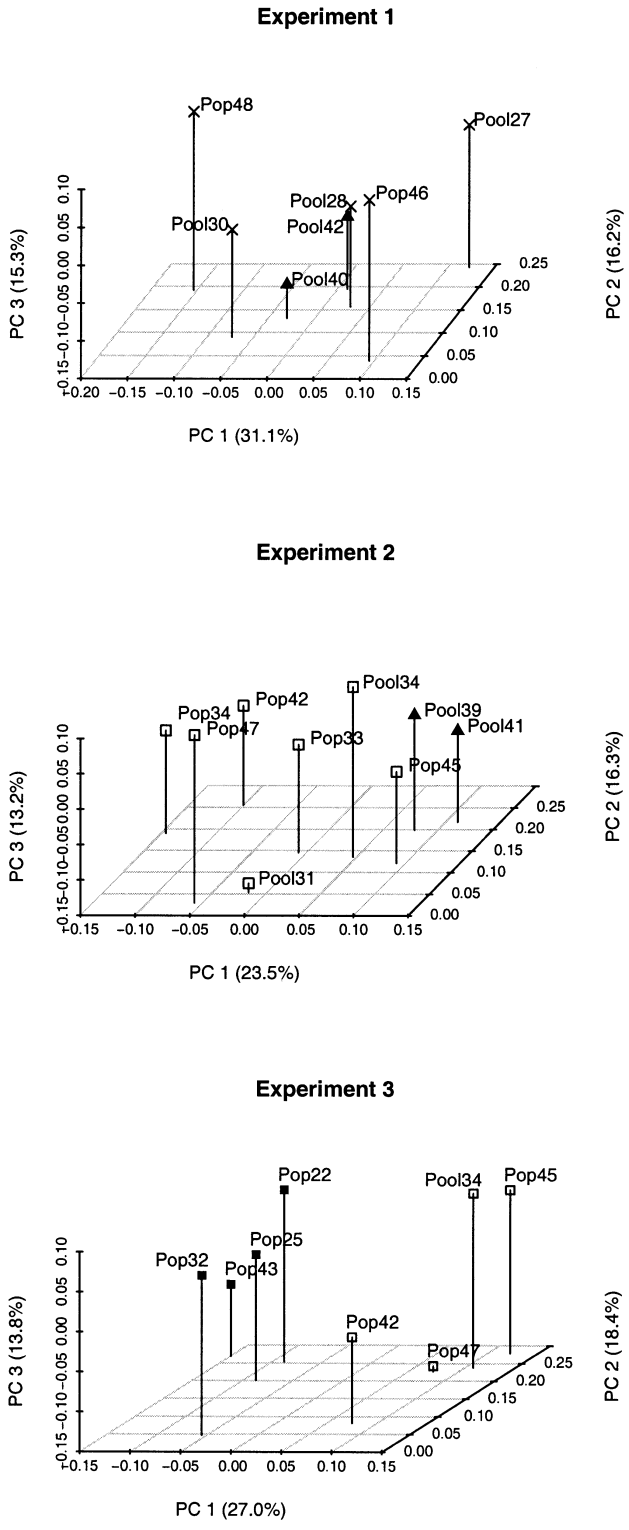

Fig. 1 Principal coordinate analysis based on the modified Roger's distance (MRD) between the populations (tropical ■, subtropical intermediate-maturity ❑, subtropical early-maturity ×, and temperate ▲ populations). PC1, PC2 and PC3 are the first, second and third principal coordinates, respectively

the high number of alleles per marker in our study, the average PIC value (0.60) was similar to those reported by Smith et al. (1997) (0.62) and Senior et al. (1998) (0.59). This can be explained by a high number of rare alleles in our study. The high within population variance revealed in the AMOVA (Table 5) can be explained by the high number of populations with a mixed origin (Table 1).

Correlation between MRD^2, $SMRD^2$ and PMPH, SCA and F_1 performance

We investigated the correlation between PMPH and MRD^2, because quantitative genetic theory suggests a linear relationship between both measures under simplifying assumptions (Falconer and Mackay 1996, pp 255). This is in harmony with related studies on mid-parent heterosis in crosses of inbred lines (see e.g., Melchinger et al. 1991; Boppenmaier et al. 1993), where the commonly employed Roger's distance is equal to MRD^2 (Melchinger 1993). A high correlation between PMPH and MRD^2 can be expected if: (1) a high association exists between heterozygosity at the marker loci and heterozygosity at quantitative trait loci (QTL), (2) heterozygosity at QTL is closely related to heterosis (Charcosset et al. 1991), (3) epistasis is absent, and (4) the populations are adapted to target environments (Moll et al. 1965).

In agreement with this expectation we found in Experiment 1 for the temperate ME a highly significant (r = 0.56**) correlation between PMPH and MRD^2 (Fig. 2). The relatively low correlation (r = 0.37) between both measures in the subtropical MEs can be explained by the non-significant SCA effects, adaption problems of crosses with the two temperate pools, and multiple alleles (Cress 1966). Nevertheless, for both MEs PMPH increased with increasing MRD^2. For Experiment 2, we observed a highly significant correlation between PMPH and MRD^2 (r = 0.53**). Here, a greater number of populations was adapted to the ME than in Experiment 1. The low correlations between both measures observed in Experiment 3 for all three MEs could be attributable to adaption problems of the parent and hybrid populations. The correlation of MRD^2 and PMPH in most experiments and MEs were higher than the correlation of MRD^2 and F_1 performance (Table 6), which is in accordance with the expectations from quantitative genetic theory (Charcosset and Essioux 1994).

To improve the low correlation of MRD^2 and SCA (Table 6), we partitioned MRD^2 into $GMRD^2$ and $SMRD^2$. Under the assumption of no epistasis and by using the parameter definitions of Gardner and Eberhart (1966), SCA can be shown to be a linear function of the $SMRD^2$ for the underlying QTL, provided all QTL have equal dominance effects (Melchinger et al. 1990). In accordance with these quantitative genetic expectations, SCA was in all instances more closely correlated with $SMRD^2$ than MRD^2 (Table 6).

The results of the first two experiments suggest that PMPH and its major component SCA increase with

954

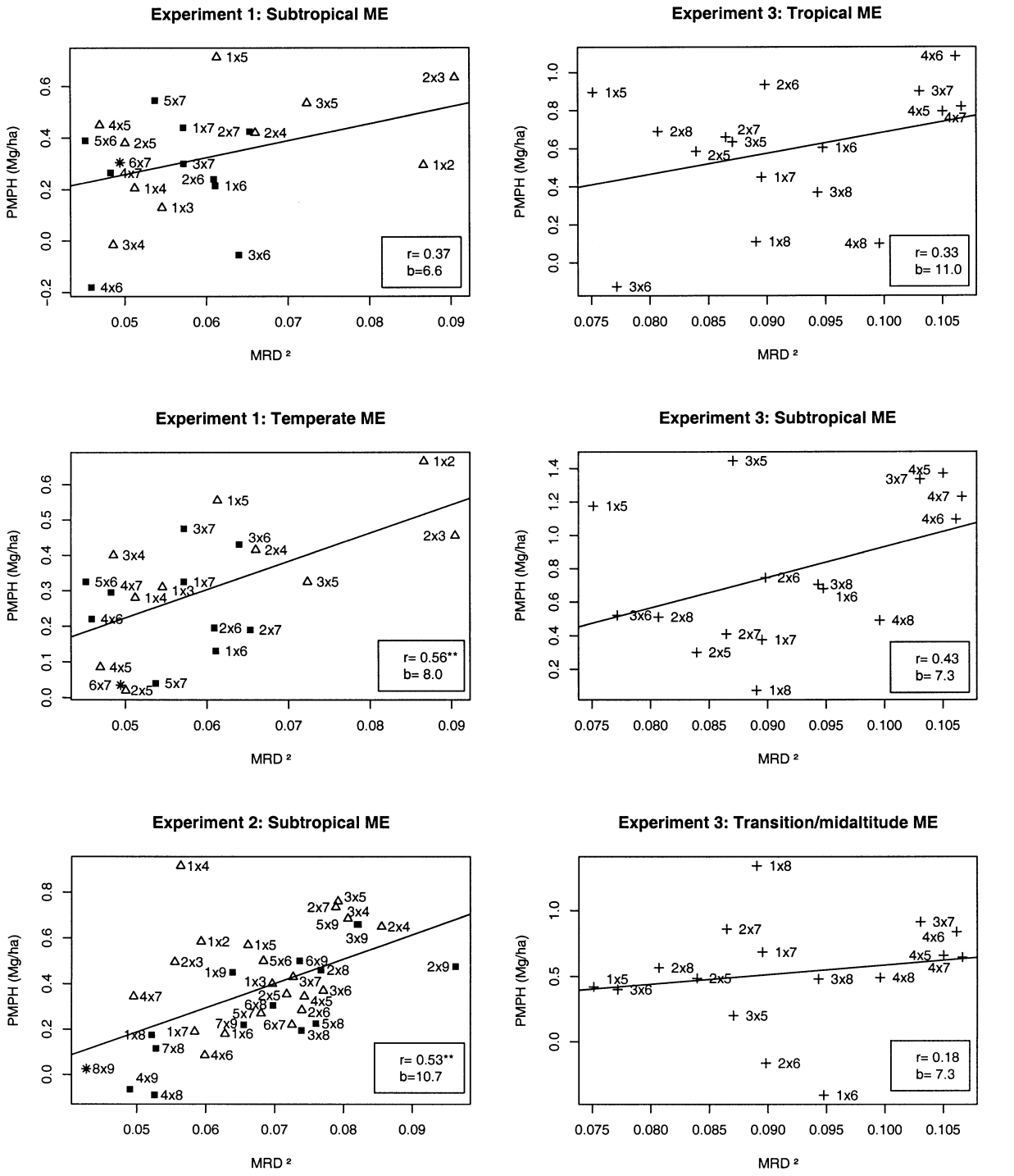

Fig. 2 Relation between squared modified Roger's distance (MRD^2) and panmictic midparent heterosis (*PMPH*) of grain yield for Experiment 1, 2, and 3 (** indicates significance at $P = 0.05$) evaluated in different mega-environments. Crosses between subtropical adapted populations ▲, between subtropical and temperate populations ■, between temperate populations *, and between subtropical × tropical populations +. Experiment1: 1 = Pop46, 2 = Pop48, 3 = Pool27, 4 = Pool28, 5 = Pool30, 6 = Pool40, 7 = Pool42; Experiment2: 1 = Pop33, 2 = Pop34, 3 = Pop42, 4 = Pop45, 5 = Pop47, 6 = Pool31, 7 = Pool34, 8 = Pool39, 9 = Pool41; Experiment3: 1 = Pop22, 2 = Pop25, 3 = Pop32, 4 = Pop43, 5 = Pool34, 6 = Pop42, 7 = Pop45, 8=Pop47

increasing genetic distance among the parent populations. Experiment 3 shows that adaption problems can cause deviations from this rule. Hence, if the populations are adapted to the target regions, genetic distance can be used as a further criterion in the search for promising heterotic patterns and groups.

Heterotic groups and patterns

Wellhausen (1978) described several heterotic patterns and identified four outstanding racial complexes: (1) Tuxpeño and related dents (Mexican, West Indian, Cuban, and Southern U.S. dents), (2) Cuban flints, (3) Coastal Tropical flints (Carribean flint), and (4) Cateto flint. He suggested to form two separate heterotic groups in the CIMMYT maize germplasm: (1) a dent composite, consisting of Tuxpeño and related dents, and (2) a flint composite consisting mainly of Cuban, Carribean and Cateto flints. However, instead of establishing two heterotic groups, CIMMYT maize breeders formed populations and pools mostly disregarding the natural heterotic patterns, which exist between the flint and dent germplasm complexes (Vasal et al. 1999), because this strategy seemed promising for breeding of OPVs. Nevertheless, some populations with a relatively pure genetic background are available (Table 1).

With the beginning of the hybrid development effort, CIMMYT conducted in the 1980s several diallel studies with different germplasm sources to detect heterotic patterns in the germplasm with mixed origin (Crossa et al.1990; Beck et al. 1991; Vasal et al. 1992a, b). Although promising heterotic patterns were suggested, it was too difficult to clearly define heterotic groups on the basis of the field data. This stimulated us to perform a combined analysis with field and molecular data for obtaining a clearer picture on promising heterotic patterns and groups.

Subtropical early-maturity germplasm

Under the preassumption of two groups, the k-means algorithm arrived for the subtropical early-maturity germplasm at the following subdivision: (1) Pool27, Pop46 and Pool28, and (2) Pool30 and Pop48. However, in the PCoA (Fig. 1) Pool28 was positioned midway between Pool27, Pop46 and Pool30, and Pop48, in accordance with pedigree information. Pop46 and Pool27 were both established using flint germplasm from the U.S., Lebanon and several European countries. Pool27 also contains white flints from Argentina. Pop48 was generated from 54 half-sib families of Pool30, which was established using dent germplasm from Europe, China, Lebanon, South America and the U.S. Cornbelt. In contrast, Pool28 was developed by mixing dent and flint germplasm from Pool30 and Pool27, respectively, which precludes their use for hybrid breeding.

Considering the field and molecular data, two heterotic groups could be formed in the subtropical early-maturity germplasm: (1) a flint composite consisting of Pop46 and Pool27, and (2) a dent composite consisting of Pop48 and Pool30.

Subtropical intermediate-maturity germplasm

With k = 3, the k-means algorithm based on MRD resulted in the following subdivision for the intermediate-maturity subtropical germplasm: (1) Pop34 and Pop42, (2) Pop33, Pop45 and Pool34, and (3) Pop47 and Pool31. These results are in accordance with the pedigree information. Pop42 and Pop34 contain ETO germplasm. The latter includes also Cuban flints and Tuxpeño germplasm. Pop33 was established using Cateto flints. Pop45 contains Cuban flints, but also Tuxpeño and a large diversity of other germplasm. Pop47 was established using 276 half-sibs of Pool32, which was established using germplasm from the same sources as Pool31. The mixed origin of Pop34, Pop45 and Pool31, Pool34 precludes their use for hybrid breeding. Hence, considering the molecular and field data two heterotic groups can be formed in the subtropical intermediate-maturity germplasm: (1) a flint composite consisting of Pop33 and Pop42, and (2) a dent composite consisting of Pop47.

In conclusion, SSR based technology offers a powerful tool for assessing the diversity among maize populations. The relationships between the populations obtained by using MRD and PCoA are in excellent agreement with the pedigree information. SSR based genetic distances in combination with field evaluation provide a solid basis for the detection of promising heterotic groups and patterns at the beginning of a hybrid breeding program.

Systematic introgression of exotic germplasm for hybrid breeding

With the increasing germplasm exchange between tropical, subtropical and temperate areas, greater options of germplasm sources are available for breeders. For hybrid breeding one has to consider the racial complexes and relationships between the populations to introgress exotic germplasm systematically in the existing heterotic groups. We investigated the use of SSR markers to achieve this goal. Considering the maturity type of the germplasm introgressed, we propose an exchange between early tropical and late subtropical, early subtropical and late temperate, germplasm and vice versa.

Pool42 was established to introduce tropical germplasm into temperate areas. The low hybrid performance of crosses with Pool42 in Experiment 1 (Table 2) suggested that it may not be of direct use for breeding programs in temperate environments. Pool39, 40 and 41 were designed to introgress temperate germplasm for the winter maize areas in the subtropics and tropics. Similar results were observed as for Pool42 (Tables 2, 3), which

indicated that they may not be valuable for breeding programs in subtropical environments. The low hybrid performance of the four pools can be explained by their low *per se* performance in all MEs.

In contrast, the high yield and PMPH of crosses between subtropical × tropical germplasm (Table 4) suggested that the exchange between both types of germplasm could benefit CIMMYT's hybrid breeding program. Aggregating all information about the relationships between the populations (Fig. 1 and Table 1) and considering the field data (Table 4), we propose an exchange of germplasm between both ETO-based Pop32 and Pop42 on one side, and the largely Tuxpeño-based Pop22 and Pop47 on the other side. Furthermore, genotypes with rare or absent SSR marker alleles in the other group and good test performance can be identified and used to systematically broaden the germplasm basis. Thus, useful alleles can be introgressed and benefit the respective breeding programs.

Acknowledgements The molecular marker analyses of this research were supported by funds from the German "Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung" Project No. 98.7860.4-001-01. Thanks to J. Crossa, G.C. Han, S. Pandey, and G. Srinivasan for providing the field data and seeds for performing this study. This paper is dedicated to Prof. Dr. h. c. F. W. Schnell on the occasion of his 90th birthday.

References

Beck DL, Vasal SK, Crossa J (1991) Heterosis and combining ability among subtropical and temperate intermediate-maturity maize germplasm. Crop Sci 31:68–73

Boppenmaier J, Melchinger AE, Seitz G, Geiger HH, Herrmann RG (1993) Genetic diversity for RFLPs in European maize inbreds. III. Performance of crosses within versus between heterotic groups for grain traits. Plant Breed 111:217–226

Brummer EC (1999) Capturing heterosis in forage crop cultivar development. Crop Sci 39:943–954

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

Charcosset A, Essioux L (1994) The effect of population structure on the relationship between heterosis and heterozygosity at marker loci. Theor Appl Genet 89:336–343

CIMMYT (1998) A complete listing of maize germplasm from CIMMYT. Maize Program Special Report, Mexico DF, Mexico

Comstock RE, Robinson HF (1948) The components of genetic variance in populations of biparental progenies and their use in estimating the average degree of dominance. Biometrics 4:254–266

Cress CE (1966) Heterosis of the hybrid related to gene-frequency differences between two populations. Genetics 53:269–274

Crossa J, Vasal SK, Beck DL (1990) Combining ability estimates of CIMMYT tropical late yellow maize germplasm. Maydica 35:273–278

Elder JK, Souther EM (1987) Computer-aided analysis of one-dimensional restriction fragments gels. In: Bishop MJ, Rawling CJ (eds) Nucleid acid and protein sequence analysis – a practical approach. IRL Press, Oxford, pp 165–172

Falconer DS, Mackay TF (1996) Introduction to quantitative genetics, 4th edn. Longman Group Ltd, London

Fisher RA (1921) On the "probable error" of a coefficient of correlation deduced from a small sample. Metron 1:1–32

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

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

Goodman MM, Stuber CW (1983) Races of maize. VI. Isozyme variation among races of maize in Bolivia. Maydica 28:169–187

Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325–338

Hallauer AR, Russell WA, Lamkey KR (1988) Corn breeding. In: Sprague GF, Dudley JW (eds) Corn and corn improvement, 3rd edn. Agron Monogr 18, ASA, CSSA and SSSA, Madison, Wisconsin, pp 463–564

Hartigan JA, Wong MA (1979) A K-means clustering algorithm. Appl Stats 28:100–108

Ihaka R, Gentleman R (1996) A language for data analysis and graphics. J Computat Graphical Stats, Vol 5 3:299–314

Lamkey KR, Edwards JW (1999) Quantitative genetics of heterosis. Chapter 10. In: Coors JG, Pandey S (eds) The genetics and exploitation of heterosis in crops. CSSA, Madison, Wisconsin

Lu H, Bernardo R (2001) Molecular marker diversity among current and historical maize inbreds. Theor Appl Genet 103:613–617

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

Melchinger AE (1999) Genetic diversity and heterosis. Chapter 10. In: Coors JG, Pandey S (eds) The genetics and exploitation of heterosis in crops. CSSA, Madison, Wisconsin

Melchinger AE, Gumber RK (1998) Overview of heterosis and heterotic groups in agronomic crops. In: Lamkey KR, Staub JE (eds) Concepts and breeding of heterosis in crop plants, CSSA, Madison, Wisconsin, pp 29–44

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

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

Michalakis Y, Excoffier L (1996) A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. Genetics 142:1061–1064

Moll RH, Salhuana WS, Robinson HF (1962) Heterosis and genetic diversity in variety crosses of maize. Crop Sci 2:197–198

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

Ron Parra J, Hallauer AR (1997) Utilization of exotic maize germplasm. Plant Breed Rev 14:165–187

Saghai-Maroof MA, Soliman KM, Jorgenson R, Allward RW (1984) Ribosomal DNA spacer length polymorphisms in barley: Mendelian inheritance, chromosomal location and population dynamics. Proc Natl Acad Sci USA 81:8014–8018

Schneider S, Roessli D, Excoffier L (2000) Arlequin, ver 2.0: a software of population genetics data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland

Senior ML, Murohy MM, Goodman MM, Stuber CW (1998) Utility of SSRs for determining genetic similarities and relationships in maize using an agarose gel system. Crop Sci 28:63–67

Smith JSC, Chin ECL, Shu H, Smith OS, Wall SJ, Senior ML, Mitchell SE, Kresovitch S, Ziegle J (1997) An evaluation of the utility of SSR loci as molecular markers in maize (*Zea mays* L.): comparisons with data from RFLPs and pedigree. Theor Appl Genet 95:163–173

Vasal SK, Srinivasan G, Crossa J, Beck DL (1992a) Heterosis and combining ability of CIMMYT's subtropical and temperate early maturity maize germplasm. Crop Sci 32:884–890

Vasal SK, Srinivasan G, Beck DL, Crossa J, Pandey S, Leon C de (1992b) Heterosis and combining ability of CIMMYT's tropical late white maize germplasm. Maydica 37:217–223

Vasal SK, Srinivasan G, Gonzalez F, Han GC, Pandey S, Beck DL, Crossa J (1992c) Heterosis and combining ability of CIMMYT's tropical × subtropical maize germplasm. Crop Sci 32:1483–1489

Vasal SK, Cordova HS, Pandey S, Srinivasan G (1999) Tropical maize and heterosis. Chapter 34. In: Coors JG, Pandey S (eds) The genetics and exploitation of heterosis in crops. CSSA, Madison, Wisconsin

Warburton ML, Xia XC, Crossa J, Franco J, Melchinger AE, Frisch M, Bohn M, Hoisington D (2002) Genetic characterisation of CIMMYT maize inbred lines and open-pollinated populations using large scale fingerprinting methods. Crop Sci (in press)

Wellhausen EJ (1978) Recent developments in maize breeding in the tropics. In: Walden DB (ed) Maize breeding and genetics. John Wiley and Sons, New York, pp 59–91

Wright S (1978) Evolution and genetics of populations, vol. IV. The University of Chicago Press

7. General Discussion

Molecular markers are a promising tool to improve the conservation of genetic diversity in seed banks and to use genetic resources for plant breeding (Brummer, 1999; Melchinger, 1999). In addition to the assessment of dissimilarities among operational taxonomic units (OTUs), the estimation of genetic diversity within OTUs is an important component to optimize the conservation and exploitation of genetic resources. Starting with the allele frequencies of the OTUs, a number of coefficients have been proposed. A thorough knowledge of the properties of these coefficients is of crucial importance to interpret the results of molecular marker-based diversity studies on a theoretically sound basis. Therefore, we characterized different coefficients for measuring diversity within and among OTUs and determined their relationships.

Genetic Diversity Measures Within OTUs

Any measure of genetic diversity within populations ought to have the following characteristics (Lewontin, 1972): (i) It should be minimum, when there is only a single allele present, because the locus shows no variation. (ii) The diversity should increase as the number of different alleles in the population increases. (iii) For a fixed number of alleles, it should be maximum when all alleles are equal in frequency. (iv) A collection of individuals made

by pooling two populations should always be equal or more diverse than the average of their individual diversities. (v) The measure should be biologically interpretable.

Several authors have used the number of alleles per locus to describe the diversity within populations (*cf.* Lu and Bernardo, 2001). The number of observed alleles is highly dependent on the sample size and, consequently, a comparison of populations with different sample sizes is not very meaningful. To overcome the problem of unequal sample sizes, Reif et al. (2004) proposed the following procedure to standardize the number of alleles: Take a random sample of individuals from each population without replacement, with the number of individuals analyzed per population being equal to the size of the smallest population. Calculate the number of alleles per population, repeat the procedure more than 10 000 times, and average across repetitions. Owing to their biological interpretation, both measures, the number of alleles and the standardized number of alleles, seem to be appealing to experimental geneticists. Nevertheless, none of the two includes information about the distribution of alleles within populations. Thus, both measures disregard the third criteria mentioned above.

The Shannon-Weaver index (H_{SW}) (Shannon and Weaver, 1949) has been developed in the context of information theory to quantify the information content of messages, but has also been widely used as a diversity measure (Lewontin, 1972):

$$H_{SW} = -K \sum_{j=1}^{n} p_j \log_b p_j, \tag{1}$$

where p_j refers to the allele frequency of the jth allele and n to the number of alleles at the locus. In contrast to the number of alleles, H_{SW} includes the distribution of the alleles. H_{SW} has been frequently applied as diversity coefficient with $K = 1$ and $b = 2$ (*cf.* Lewontin, 1972) or $b = e$ (*cf.* Grenier et al., 2001), although it is not biologically interpretable.

The effective number of alleles (n_e) was defined by Kimura and Crow

(1964) as the reciprocal of homozygosity:

$$n_e = \frac{1}{\sum_{j=1}^{n} p_j^2}. \tag{2}$$

The measure n_e is equal to the actual number of alleles if and only if all alleles have the same frequency, otherwise it is smaller. Similarly to H_{SW}, n_e is not biologically interpretable.

An alternative genetic diversity coefficient, sometimes referred to as average heterozygosity, but more precisely described by the term gene diversity (H_S) was suggested by Nei (1987):

$$H_S = 1 - \sum_{j=1}^{n} p_j^2 = 1 - \frac{1}{n_e}. \tag{3}$$

In a random mating population, H_S can be considered as the average proportion of heterozygotes per locus. For diploid populations not in HWE or for haploid organisms, H_S does not equal the average proportion of heterozygotes. In this case, H_S can be interpreted as the probability that two randomly chosen individuals have different alleles at the locus under consideration. H_S fulfills all five above-mentioned criteria. In consequence, H_S can be recommended for many applications in plant breeding and seed bank management.

An alternative diversity measure is the average of all pairwise genetic distances among individuals within a population ($\bar{d}$). It is based on the allele frequencies of the individuals of the population and not on the population allele frequencies. Consider one locus and n homozygous inbred lines extracted from a population. Assuming that (a) $\bar{d}$ is based on the modified Rogers distance (d_W) and (b) r_i denotes the number of individuals carrying

the allele A_i, it implies that:

$$\begin{aligned}
\bar{d_W} &= \frac{1}{\binom{n}{2}} \sum_{k=1}^{n} \sum_{j=k+1}^{n} d_W(j,k) \\
&= 1 - \frac{1}{\binom{n}{2}} \left(\binom{r_1}{2} + \binom{r_2}{2} + \cdots + \binom{r_i}{2} \right) \\
&= 1 - \left(\frac{r_1(r_1-1) + r_2(r_2-1) + \cdots + r_i(r_i-1)}{n(n-1)} \right) \\
&= 1 - \frac{1}{n(n-1)} (\sum_{i=1}^{n} r_i^2 - \sum_{i=1}^{n} r_i)
\end{aligned}$$

using $p_i = \frac{r_i}{n}$ and $\sum_{i=1}^{n} r_i = n$

$$\begin{aligned}
\bar{d_W} &= 1 - \frac{n}{n-1} \sum_{i=1}^{n} p_i^2 + \frac{1}{n-1} \\
&= 1 - \left(\frac{n-1+1}{n-1} \right) \sum_{i=1}^{n} p_i^2 + \frac{1}{n-1} \\
&= 1 - \sum_{i=1}^{n} p_i^2 + \frac{1}{n-1} (1 - \sum_{i=1}^{n} p_i^2) \\
&= \frac{n}{n-1} (1 - \sum_{i=1}^{n} p_i^2) \\
&= \frac{n}{n-1} H_S
\end{aligned}$$

Under the above-mentioned assumption, H_S is related to $\bar{d_W}$ and fulfills the five criteria of Lewontin (1972). The measure $\bar{d_W}$ is, therefore, a suitable criterion for measuring diversity within OTUs.

Genetic Dissimilarity Measures Between OTUs

Assume that H_i and H_j denote measures to determine genetic diversity within populations i and j, respectively, and that H_{ij} refers to genetic diversity within the pooled population of i and j. A diversity measure between

the two populations i and j (d_{ij}) can then be defined as (Rao, 1982):

$$d_{ij} = H_{ij} - 1/2(H_i + H_j). \tag{4}$$

This diversity measure among populations can also be interpreted as a dissimilarity measure between populations. Nei (1975) showed for $H_{ij} = -\ln(1 - H_{Sij})$, $H_i = -\ln(1 - H_{Si})$, and $H_j = -\ln(1 - H_{Sj})$ that:

$$d_{ij} = -\ln \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} p_{ij} q_{ij}}{\sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n_i} p_{ij}^2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} q_{ij}^2}} = d_{N72}. \tag{5}$$

Under those conditions d_{ij} is called Nei's standard genetic dissimilarity (d_{N72}). The coefficient d_{N72} was developed assuming (i) the infinite-allele model (Kimura and Crow, 1964) and (ii) that an ancestral population was split into several subpopulations of equal sizes that subsequently diverged due to drift and mutation. If (a) the mutation-drift balance is maintained throughout the evolutionary process, (b) selection is absent, and (c) the dissimilarity is not very large, then the expected value of d_{N72} is proportional to the time since the subpopulations diverged (Nei et al., 1983).

Both dissimilarity coefficients between OTUs, Reynolds dissimilarity (d_{RE}) and Cavalli-Sforza and Edward's distance (d_{CE}), are based on similar evolutionary models as d_{N72}, but differ in the force causing divergence between subpopulations. The coefficient d_{RE} was developed assuming that an ancestral population was split into several subpopulations of identical sizes that subsequently diverged due to drift. The coefficient d_{CE} was developed based on Kimura's (1954) model of "selective drift" by assuming that (i) the mutation rate is small and (ii) variation in selection pressure is rapid and haphazard. All three above-mentioned dissimilarity measures are based on evolutionary models, whose underlying assumptions are presumably not met in seed bank and plant breeding materials. Nevertheless, d_{CE}, d_{RE}, and d_{N72} have been frequently applied in surveys with breeding and seed bank germplasm (*e.g.*, DeHaan et al., 2003; Labate et al., 2003).

The modified Rogers (d_W) and the Rogers distance (d_R) are both modifications of the Euclidean distance and possess interesting genetical properties. The coefficient d_R is linearly related to the coefficient of coancestry

(Melchinger et al., 1991), which makes it appropriate to uncover pedigree relationships in plant breeding or seed bank material. Consequently, d_R was used in our survey investigating a possible genetic diversity loss during the domestication and breeding of wheat. Melchinger (1999) showed that d_W^2 is linearly related to the panmictic-midparent heterosis and was therefore appropriate to examine (i) the prediction of heterosis with genetic distances or (ii) the establishment of heterotic groups in seed bank and plant breeding material. Therefore, we applied d_W^2 in our study with CIMMYT maize germplasm.

Flux of Diversity in Wheat

Bread wheat was domesticated 10 000 years ago in the Fertile Crescent (Salamini et al., 2002). It is postulated that the size of the founder population of bread wheat was limited, causing a domestication bottleneck (Fig. 1). We observed a non-significant decrease in standardized number of alleles (N_a) and gene diversity (H) from *T. tauschii* accessions to landrace cultivars, resulting in a significant ($P < 0.1$) diversity loss ($\Delta H = 0.19$). These results, together with the findings of 2.5 unique alleles per locus present in *T. tauschii* but not in landrace cultivars, indicate a reduction in genetic variation during the process of wheat domestication. This is in agreement with previous studies reporting that the *T. tauschii* genome contains considerably more genetic variation than the D genome of hexaploid wheat (Lubbers et al., 1991; Lelley et al., 2000). The reduction in genetic diversity is probably the product of a relatively young history of the wheat crop, a presumably small founder population, and an intensive long-term selection for agronomic traits.

During the past century, traditional landrace cultivars were continuously replaced by modern wheat cultivars, which culminated in only about 3% of the wheat growing area currently sown with landrace cultivars (Smale et al.,

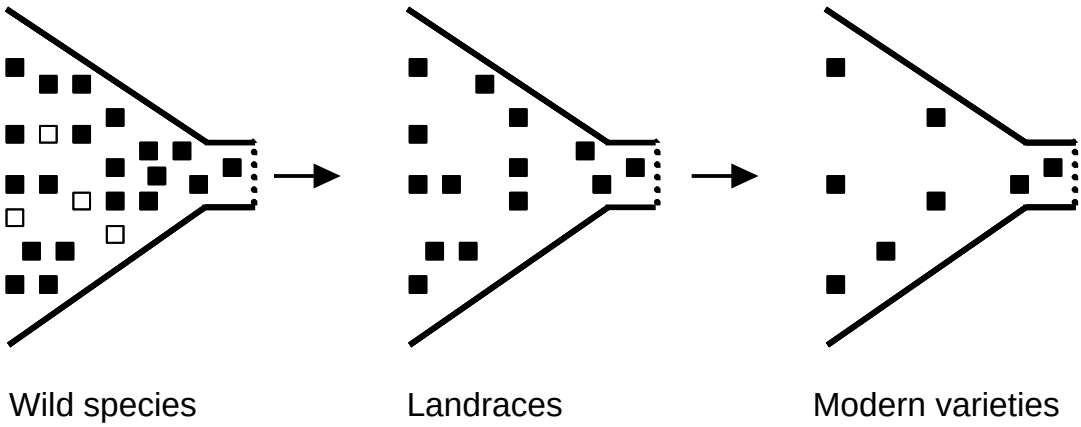

Figure 1. Genetic bottlenecks imposed on crop plants during domestication and through modern plant-breeding practices. Boxes represent allelic variation of genes originally found in the wild, but gradually lost during domestication and breeding (adapted from Tanksley and McCouch, 1997)

2002). Modern wheat cultivars were bred with a limited number of landrace cultivars in their pedigree and it is postulated that modern wheat cultivars contain less genetic diversity than landrace cultivars (Frankel, 1970). Combining all SSRs, a loss of gene diversity ΔH of 0.05 was revealed from landrace cultivars to modern wheat cultivars. Together with the observation that 1.9 unique alleles per locus were present in landrace cultivars but absent in modern wheat cultivars, this indicated a substantial genetic diversity loss from landrace cultivars to modern wheat cultivars. This outcome can be explained by (i) the limited number of landrace cultivars used as the germplasm base for the development of modern wheat cultivars and (ii) selection and drift during the breeding of modern wheat cultivars. The loss of genetic diversity may indicate an elimination of undesired or even deleterious alleles, but may also reflect an erosion of alleles valuable for plant improvement and future demands of producers and consumers. The latter hypothesis was supported by various surveys reporting the potential of landrace cultivars as a source of novel useful allelic variation (Cox et al., 1992; Villareal et al., 1995).

It has been claimed that plant breeding reduces genetic diversity in elite germplasm, which could jeopardize future selection gain in crop improve-

ment (Tanksley and McCouch, 1997). Genetic diversity, measured as average Rogers distances between individuals, was narrowed from 1950 to 1989, but was enhanced from 1990 to 1997. Our results indicate that breeders averted the narrowing of the germplasm base and subsequently increased the genetic diversity through the introgression of various novel wheat materials.

Summarizing, we observed a diversity loss from *T. tauschii* to landrace cultivars and from landrace cultivars to modern wheat cultivars. Consequently, both landrace cultivars and *T. tauschii* represent useful sources for broadening the genetic base of elite wheat breeding germplasm. Favorable alleles from the landrace cultivars can be introgressed into the elite germplasm pool with classical breeding methods. The introgression of *T. tauschii* alleles can be achieved via the creation of synthetic hexaploid wheats.

Conservation of CIMMYT's Maize Germplasm

From 1964 until 1973, CIMMYT developed and improved a wide array of maize germplasm. Each population was established with materials from a single racial complex. In 1974, a major shift in the organization of the germplasm was initiated. CIMMYT devoted its efforts to the formation and development of broad-based populations and pools, mostly disregarding the racial complexes. According to its adaptation, this germplasm was grouped into four mega-environments (MEs) and subsequently improved through recurrent intra-population selection (Vasal et al., 1999). Principal coordinate analysis based on allele frequencies of the populations revealed that germplasm adapted to the same ME clustered together. Thus, the grouping of the maize populations into MEs was clearly supported by the SSR data.

The analysis of 23 maize populations also revealed that most of the genetic diversity was within the populations and just a minor part between the

populations. This reflects CIMMYT's breeding policy and the establishment of the germplasm. Most maize populations were composed of germplasm from several racial complexes and have been improved with intra-population breeding methods. Our results indicate that the applied procedures to handle the broad range of germplasm were suboptimal with regard to (i) maintaining maximum genetic diversity within the populations and (ii) conserving genetic diversity between the populations. It is rather likely that desired alleles, which occurred with high frequency in just one racial complex, are lost by mixing different germplasm sources. Consequently, the conservation of the genetic diversity within and among CIMMYT's maize germplasm can be optimized by considering the racial complexes of the populations.

Use of CIMMYT's Maize Germplasm for Plant Breeding

The improved populations of CIMMYT have played an important role in maize production in developing countries as open pollinated varieties (OPV) (Vasal et al., 1999). With the decision in 1984 to initiate a hybrid breeding program, CIMMYT conducted several diallel studies to identify heterotic groups and patterns among the populations (Beck et al., 1991; Crossa et al., 1990; Vasal et al., 1992a,b). Some promising heterotic patterns were suggested, but a clear grouping of the maize germplasm on the basis of the field data was difficult due to the partially mixed origin of the populations.

If the panmictic midparent heterosis (PMPH) increases with increasing genetic distances of the parents, molecular marker-based genetic distances are a valuable tool to establish heterotic groups and patterns in conjunction with field trials (Melchinger, 1999). The results of our study indicate that PMPH increases with increasing genetic distance among the parent populations and that adaption problems can cause deviations from this rule. Hence, if the populations are adapted to the target regions, genetic distance can be used as

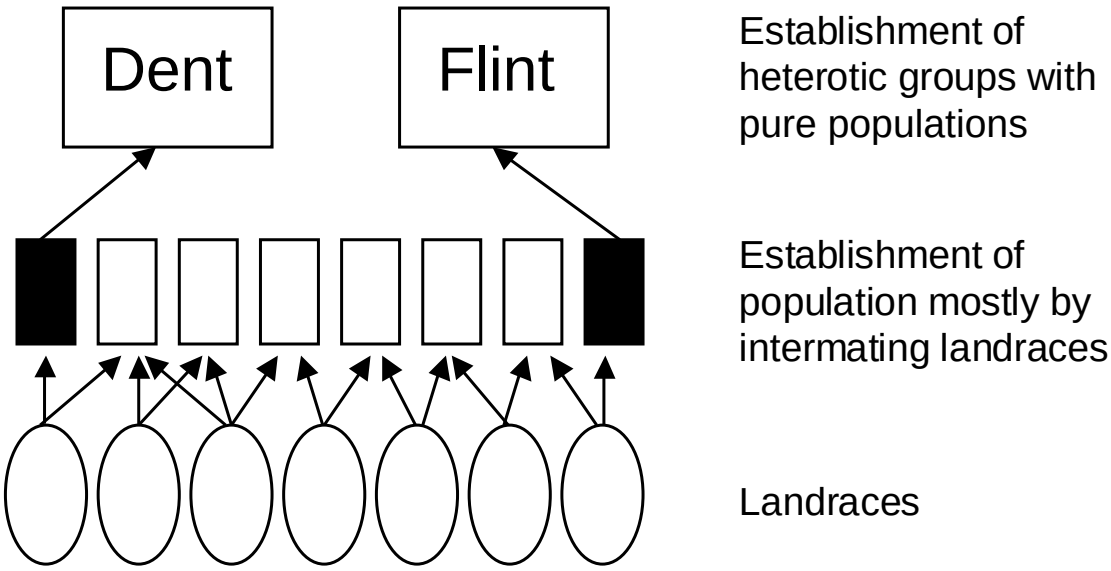

Figure 2. Establishment of heterotic groups in CIMMYT's maize germplasm.

a further criterion in the search for promising heterotic groups and patterns in addition to field trials.

Germplasm based on different racial complexes is useful for the improvement of OPVs. Nevertheless, the significant positive correlation between genetic distances and heterosis supports the concept of clearly distinct heterotic groups. Populations with a mixed constitution are therefore not suitable for hybrid breeding. The reduced genetic diversity among the populations caused by admixture can only be recovered by long-term isolation or reciprocal recurrent selection programs. Therefore, only few CIMMYT populations based on one racial complex (P21, P32, P33, P42, P43, and Pl24) are suitable for hybrid breeding (Fig. 2). For all MEs, a dent $\times$ flint heterotic pattern seems to be most promising. As the number of 'pure' flint and dent populations is limited, their genetic base can be broadened by introgression of (i) 'pure populations' conserved since the admixture of 1973, (ii) 'pure germplasm' from other MEs, and (iii) landraces stored in the seed bank.

By focusing on a limited number of available populations in the hybrid program, the question arises whether functional genetic diversity can be lost

during the transition from OPVs to hybrids. Association mapping was proposed as a promising method to detect genes and alleles of interest (Lynch and Walsh, 1997). Populations not used for hybrid breeding can therefore be systematically mined for favorable alleles, once a gene of agronomic importance is detected.

The resolution of association studies in a sample depends on the extent of linkage disequilibrium (LD) across the genome. We found that less than 0.3% of the two-locus disequilibrium tests were significant. On one hand, the lack in LD in our study can be explained by the low-density marker map and the decrease of LD with successive generations of intermating since the establishment of the populations. On the other hand, the sample size of 48 individuals per population, the precision in estimating haplotype frequencies with the EM algorithm (Excoffier and Slatkin, 1995), and the elimination of loci deviating from HWE result in a low power to detect LD. Our results indicate that difficulties caused by non-detection of heterozygous individuals should be avoided by fingerprinting inbred lines extracted from the populations. Thus, the precision of LD detection can be enhanced. Detailed investigations with denser marker maps are required to determine the resolution of association studies in this germplasm.

Strategy to Optimize the Exploitation of CIMMYT's Maize Genetic Resources

A systematic exploitation of CIMMYT's maize genetic resources for breeding is lacking. A possible strategy could rest upon the natural relationships between various races of maize. Goodman and Brown (1988) summarized the available information about the classification of races from the Western Hemisphere. This grouping of the landraces is only based on phenotypic and chromosome knob data. Further detailed investigations based

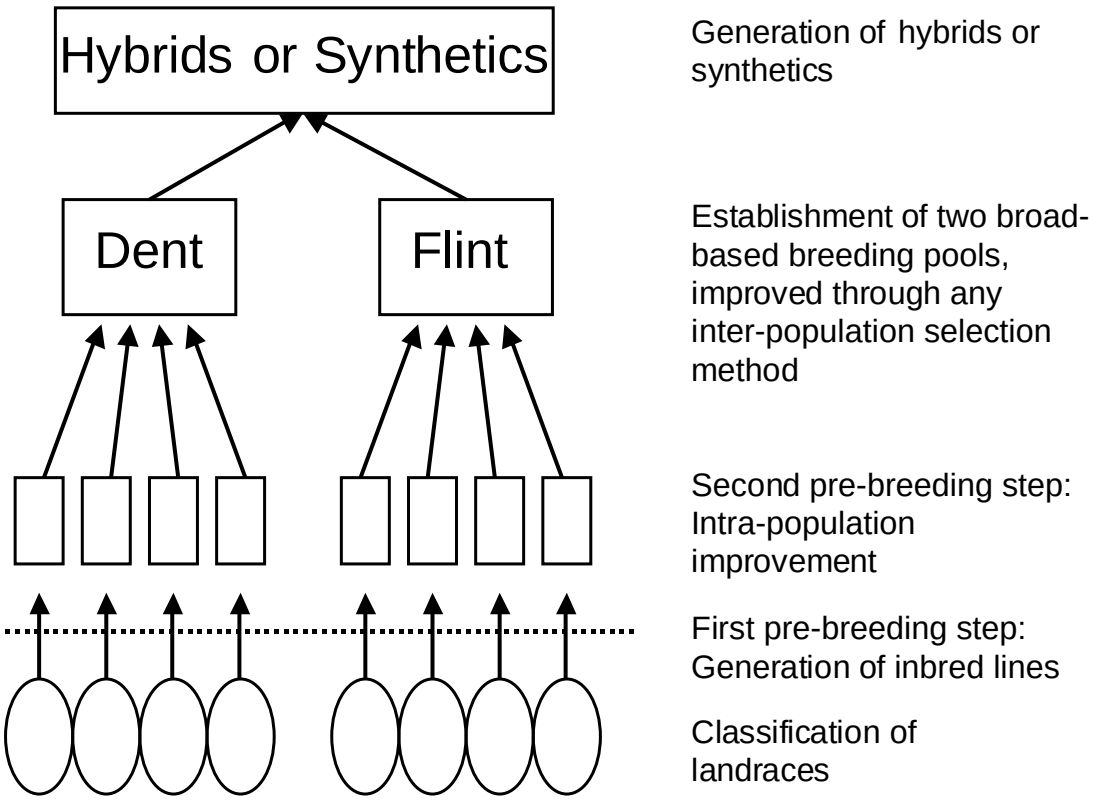

Figure 3. Strategy to optimize the exploitation of CIMMYT's maize genetic resources for breeding.

on molecular markers are required to obtain precise information about the relationships among the diverse races.

After having the races classified, they could be evaluated in field trials and first pre-breeding steps should be undertaken. One possible pre-breeding step consists of the decomposition of each race into a representative sample of homozygous inbred lines (Fig. 3). The decomposition has two advantages: (i) the obtained genotypes are fixed and, consequently, phenotypic information can be collected for a specific genotype, which is for example important for association mapping, and (ii) deleterious alleles are eliminated during the inbreeding process. The inbred lines can either be generated by selfing a sample of selected genotypes or by using the newly emerging technique of producing doubled-haploids (DHs) in maize (Eder and Chalyk, 2002). DHs

have the advantage of a short time required to obtain fully inbred lines. As a second pre-breeding step, any intra-population selection method could be applied to enhance the agronomic performance of the inbred lines.

Based on currently available data, it can be recommended to establish two broad-based breeding pools: (i) a dent and (ii) a flint composite, as already suggested in an assay of Wellhausen (1978). Each of these composites can be further divided into subpopulations for each ME. For the short range, it might be best to start with the more narrowly based 'pure' breeding populations P21, P32, P33, P42, P43, and Pl24 to establish the two breeding pools. Once the pools are established, any method of reciprocal recurrent selection could be applied to enhance their combining ability. The genetic base of the breeding pools can subsequently be increased by introgression of inbred lines of the various races.

Where OPVs are desired, superior inbred lines from both breeding pools could be combined in a synthetic and further improved through intra-population improvement in the final target environment. The combination of the flint and dent germplasm pools into a single population should provide the best base to obtain a synthetic with high yielding performance. However, hybrids present the most efficient way of exploiting the strong natural heterotic pattern existing between the dent and flint complexes (Duvick, 2001). Thus, if feasible and practical, hybrids should be generated by crosses between opposite heterotic breeding pools.

The suggested strategy would guarantee (i) an optimal use of the promising dent $\times$ flint heterotic pattern for marginal environments via improved synthetics as well as for high-input environments via hybrids and (ii) a systematic exploitation of the available genetic diversity in maize.

References

Beck, D.L., S.K. Vasal, and J. Crossa. 1991. Heterosis and combining ability among subtropical and temperate intermediate-maturity maize germplasm. Crop Sci. 31:68–73.

Brummer, E.C. 1999. Capturing heterosis in forage crop cultivar development. Crop Sci. 39:943–954.

Cox, T.S., W.J. Wilson, D.L. Gill, S. Leath, W.W. Bockus, and L.E. Browder. 1992. Resistance to foliar diseases in a collection of *T. tauschii* germplasm. Plant Disease 76:1061–1064.

Crossa, J., S.K. Vasal, and D.L. Beck. 1990. Combining ability estimates of CIMMYT tropical late yellow maize germplasm. Maydica 35:273–278.

DeHaan, L.R., N.J. Ehlke, C.C. Sheaffer, G.J. Muehlbauer, and D.L. Wyse. 2003. Illinois bundleflower genetic diversity determined by AFLP analysis. Crop Sci. 43:402–408.

Duvick, D.N. 2001. Biotechnology in the 1930s: the developement of hybrid maize. Nat. Rev. Genet. 2:69–74.

Eder, J., and S. Chalyk. 2002. In vivo haploid induction in maize. Theor. Appl. Genet. 104:703–708.

Excoffier, L., and M. Slatkin. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol. Biol. Evol. 12:921–927.

Frankel, O.H. 1970. Genetic dangers of the Green Revolution. World Agric. 19:9–14.

Goodman, M.M., and W.L. Brown. 1988. Corn breeding. p. 33–79. *In* G.F. Sprague and J.W. Dudley (eds.) Corn and Corn Improvement. 3rd ed. Agron. Monogr. 18. ASA, CSSA, and SSSA, Madison, WI.

Grenier, C., P. Hamon, and P.J. Bramel-Cox. 2001. Core collection of sorghum II. Comparison of three random sampling strategies. Crop Sci. 41:241–246.

Kimura, M. 1954. Process leading to quasi-fixation of genes in natural populations due to random fluctuation of selection intensities. Genetics 39:280–295.

Kimura, M., and J.F. Crow. 1964. The number of alleles that can be maintained in a finite population. Genetics 49:725–738.

Labate, J.A., K.R. Lamkey, S.H. Mitchell, S. Kresovich, H. Sullivan, and J.S.C. Smith. 2003. Molecular and historical aspects of Corn Belt dent diversity. Crop Sci. 43:80–91.

Lelley, T., M. Stachel, H. Grausgruber, and J. Vollmann. 2000. Analysis of relationships between *Aegilops tauschii* and the D genome of wheat utilizing microsatellites. Genome 43:661–668.

Lewontin, R.C. 1972. The apportionment of human diversity. Evolutionary Biology 6:381–398.

Lu, H., and R. Bernardo. 2001. Molecular marker diversity among current and historical maize inbreds. Theor. Appl. Genet. 103:613–617.

Lubbers, E.L., K.S. Gill, T.S. Cox, and B.S. Gill. 1991. Variation of molecular markers among geographically diverse accessions of *Triticum tauschii*. Genome 34:354–361.

Lynch, M., and B. Walsh. 1997. Genetics and Analysis of Quantitative Traits. p. 413. Sinauer Assoc., Sunderland, MA.

Melchinger, A.E. 1999. Genetic diversity and heterosis. p. 99–118. *In* J.G. Coors and S. Pandey (eds.) The Genetics and Exploitation of Heterosis in Crops. CSSA, Madison, WI.

Melchinger, A.E., M.M. Messmer, M. Lee, W.L. Woodman, and K.R. Lamkey. 1991. Diversity and relationships among U.S. maize inbreds revealed by restriction fragment length polymorphisms. Crop Sci. 31:669–678.

Nei, M. 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

Nei, M. 1975. Molecular Population Genetics and Evolution. North-Holland, Amsterdam and New York.

Nei, M., F. Tatjima, and Y. Tateno. 1983. Accuracy of estimated phylogenetic trees from molecular data. J. Mol. Evol. 19:153–170.

Rao, C.R. 1982. Diversity and dissimilarity coefficients: a unified approach. Theoretical Population Biology 21:24–43.

Reif, J.C., X.C. Xia, A.E. Melchinger, M.L. Warburton, D.A. Hoisington, D. Beck, M. Bohn, and M. Frisch. 2004. Genetic diversity determined within and among CIMMYT maize populations of tropical, subtropical, and temperate germplasm by SSR markers. Crop Sci. 44:326–334.

Salamini, F., H. Özkan, A. Brandolini, R. Schäfer-Pregl, and W. Martin. 2002. Genetics and geography of wild cereal domestication in the near east. Nat. Rev. Genet. 3:429–441.

Shannon, C.E., and W. Weaver. 1949. The Mathematical Theory of Communication. Univ. Illinois Press, Urbana.

Smale, M., M.P. Reynolds, M.L. Warburton, B. Skovmand, R. Trethowan, R.P. Singh, I. Ortiz-Monasterio, and J. Crossa. 2002. Dimension of diversity in modern spring bread wheat in developing countries from 1965. Crop Sci. 42:1766–1779.

Tanksley, S.D., and S.R. McCouch. 1997. Seed banks and molecular maps: unlocking genetic potential from the wild. Science 277:1063–1066.

Vasal, S.K., G. Srinivasan, D.L. Beck, J. Crossa, S. Pandey, and C. de Leon. 1992a. Heterosis and combining ability of CIMMYT's tropical late white maize germplasm. Maydica 37:217–223.

Vasal, S.K., G. Srinivasan, J. Crossa, and D.L. Beck. 1992b. Heterosis and combining ability of CIMMYT's subtropical and temperate early maturity maize germplasm. Crop Sci. 32:884–890.

Vasal, S.K., H.S. Cordova, S. Pandey, and G. Srinivasan. 1999. Tropical maize and heterosis. p. 363–373. *In* J.G. Coors and S. Pandey (eds.) The Genetics and Exploitation of Heterosis in Crops. CSSA, Madison, WI.

Villareal, R.L., G.F. Davila, and A.M. Kazi. 1995. Synthetic hexaploids × *Triticum aestivum* advanced derivatives resistant to Karnal Bunt (*Tilletia indica Mitra*). Cereal Res. Com. 27:127–132.

Wellhausen, E.J. 1978. Recent developments in maize breeding in the tropics. p. 59–91. *In* D.B. Walden (ed.) Maize Breeding and Genetics. John Wiley & Sons, New York.

8. Summary

Genetic diversity is a valuable natural resource and plays a key role in future breeding progress. Germplasm collections as a source of genetic diversity must be well-characterized for an efficient management and effective exploitation. The advent of PCR-based molecular markers such as simple sequence repeats (SSRs) has created an opportunity for fine-scale genetic characterization of germplasm collections. The objective of this research was to optimize the utilization of genetic resources conserved at the International Wheat and Maize Improvement Center (CIMMYT), with the aid of DNA markers.

Choice of suitable dissimilarity measures is important to facilitate the interpretation of findings from DNA marker studies on a theoretically sound basis. The objective of a theoretical study was to examine 10 dissimilarity coefficients widely used in germplasm surveys, with special focus on applications in plant breeding and seed banks. The distance and Euclidean properties of the dissimilarity coefficients were investigated as well as the underlying genetic models. Application areas for different coefficients were suggested on the basis of the theoretical findings.

It has been claimed that plant breeding reduces genetic diversity in elite germplasm, which could seriously jeopardize the continued ability to improve crops. The objectives of the presented experimental study with wheat were to examine the loss of genetic diversity during (i) domestication of the species,

(ii) change from traditional landrace cultivars (LC) to modern breeding varieties, and (iii) intensive selection over 50 years of international breeding. A sample of 253 CIMMYT or CIMMYT-related modern wheat cultivars, LC, and *Triticum tauschii* accessions were characterized with up to 90 SSR markers covering the entire wheat genome.

A loss of genetic diversity was observed from *T. tauschii* to LC and from LC to the elite breeding germplasm. Wheat genetic diversity was narrowed from 1950 to 1989, but was enhanced from 1990 to 1997. The results indicate that breeders averted the narrowing of the wheat germplasm base and subsequently increased the genetic diversity through the introgression of novel materials. The LC and *T. tauschii* contain numerous unique alleles that were absent in modern wheat cultivars. Consequently, both LC and *T. tauschii* represent useful sources for broadening the genetic base of elite wheat breeding germplasm.

In the 1980's, CIMMYT generated more than 100 maize populations and pools but little is known about the genetic diversity of this germplasm. The objective of the study with 23 CIMMYT maize populations was to characterize their population genetic structure with SSRs. The populations adapted to tropical, subtropical intermediate-maturity, subtropical early-maturity, and temperate mega-environments (ME) were fingerprinted with 83 SSR markers. Estimates of genetic differentiation (G_{ST}) between populations revealed that most of the molecular variation was found within the populations. Principal coordinate analysis based on allele frequencies of the populations revealed that populations adapted to the same ME clustered together and, thus, supported clearly the ME structure. Novel strategies were suggested to optimize the conservation of the genetic diversity within and among the populations.

Heterotic groups and patterns are of fundamental importance in hybrid breeding. The objective of the presented study with a subset of 20 out of the 23 maize populations was to investigate the relationship between heterosis and genetic distance determined with SSR markers. The published data of

three diallels and one factorial trial evaluated for grain yield were re-analyzed to calculate heterosis in population hybrids. Correlations of squared modified Rogers distance and heterosis were mostly positive and significant, but adaption problems caused deviations in some cases. For populations adapted to the target regions, genetic distance can be used as a further criterion in the search for promising heterotic patterns and groups. For intermediate- and early-maturity subtropical germplasm, two heterotic groups were suggested, consisting of a flint and dent composite. For the tropical germplasm, it was possible to assign population (Pop29) to the established heterotic group A and propose new heterotic groups (Pop25, Pop43).

Our experimental results corroborate that SSRs are a powerful tool to (i) detect relationships among different germplasm, (ii) assess the level of genetic diversity present in germplasm pools and its flux over time, and (iii) search for promising heterotic groups for hybrid breeding in complementation to field trials.

9. Zusammenfassung

Die genetische Diversität ist für den zukünftigen Züchtungsfortschritt von zentraler Bedeutung. In Genbanken ist ein bedeutender Anteil der Diversität von Nahrungspflanzen konserviert. Eine optimale Erhaltung und bestmögliche Nutzung dieser genetischen Vielfalt bedarf einer fundierten Charakterisierung der vorhandenen Genotypen. DNA Marker stellen hierzu ein vielversprechendes Werkzeug dar. Die vorliegende Arbeit befasst sich daher mit dem Einsatz von Markertechnologie zur Nutzbarmachung genetischer Ressourcen des Internationalen Mais- und Weizenforschungszentrums (CIMMYT) für die Pflanzenzüchtung.

Die Wahl eines geeigneten Ähnlichkeitskoeffizienten spielt bei der Interpretation von Ergebnissen aus DNA-Markerstudien eine entscheidende Rolle. In einer theoretischen Untersuchung wurden zehn häufig in Diversitätsanalysen benutzte Ähnlichkeitskoeffizienten im Hinblick auf ihre Eignung für Pflanzenzüchtungs- und Genbankstudien untersucht. Die den Ähnlichkeitskoeffizienten zugrundeliegenden mathematischen und genetischen Konzepte wurden detailliert diskutiert. Auf der Grundlage dieser Ergebnisse konnten für die unterschiedlichen Koeffizienten Anwendungsgebiete vorgeschlagen werden.

Eine populäre Hypothese ist, dass Pflanzenzüchtung die genetische Diversität im Elitezuchtmaterial reduziert und somit den zukünftigen Zuchtfortschritt gefährdet. Ziel einer experimentellen Arbeit mit Weizen war, einen möglichen Diversitätsverlust zu untersuchen während (i) der Domestikation

dieser Art, (ii) dem Übergang von traditionellen Landsorten (LC) zu modernen Weizensorten (MWC) und (iii) 50 Jahren intensiver Selektion durch internationale Weizenzüchtung. Eine Stichprobe von 253 CIMMYT oder CIMMYT verwandten MWC, LC und *Triticum tauschii* Akzessionen wurde mit 90 SSRs genotypisiert.

Ein drastischer genetischer Diversitätverlust wurde beim Vergleich von *T. tauschii* mit den LR und LR mit den MWC beobachtet. Die genetische Vielfalt von MWC nahm von 1950 bis 1989 ab, stieg aber von 1990 bis 1997 wieder an. Die Befunde deuten darauf hin, dass die Weizenzüchter am CIMMYT die Gefahr einer Einengung der genetischen Basis erkannten und erfolgreich die genetische Diversität im Zuchtmaterial durch Introgression neuer Genressourcen erweiterten. Zahlreiche Allele waren in LC oder in *T. tauschii* vorhanden, die jedoch in MWC nicht gefunden wurden. Folglich stellen sowohl LC als auch *T. tauschii* eine wertvolle Ressource zur Erweiterung der genetischen Basis des Elitezuchtmaterials bei Weizen dar.

In den 80'er Jahren wurden am CIMMYT über 100 Maispopulationen etabliert. Allerdings ist wenig über die genetische Diversität dieses Pflanzenmaterials bekannt. Eine Untersuchung von 23 Maispopulationen zielte auf die Charakterisierung ihrer populationsgenetischen Struktur mit SSR Marker Daten ab. Insgesamt 672 Genotypen der Maispopulationen, adaptiert an tropische, subtropische und gemäßigte Anbauzonen (ME), wurden mittels 83 SSR Markern molekularbiologisch charakterisiert. Der größte Teil der genetischen Varianz wurde innerhalb der Maispopulationen detektiert und der geringere Teil zwischen den Populationen. Eine Hauptkoordinatenanalyse, basierend auf den Populationsallelfrequenzen, ergab eine Gruppierung von Populationen, die an die gleichen Umweltbedingungen adaptiert sind und stützt somit die Einteilung in ME. Es konnten alternative Strategien vorgeschlagen werden, um den Erhalt der genetischen Diversität zwischen und innerhalb der Populationen zu verbessern.

Heterotische Gruppen sind von grundlegender Bedeutung in der Hybridzüchtung. Eine Studie mit 20 der 23 Maispopulationen sollte die Beziehung

zwischen Heterosis und genetischen Distanzen auf der Grundlage von SSR Markern untersuchen. Publizierte Ergebnisse für den Kornertrag von vier Experimenten mit diallelen bzw. faktoriellen Populationskreuzungen wurden reanalysiert und der Heterosiszuwachs der Populationshybriden berechnet. Die Korrelationen zwischen genetischen Distanzen und Heterosiszuwachs waren meist positiv und signifikant. Allerdings verursachten Adaptionsprobleme in einigen Fällen Abweichungen. Bei Populationen, die an die Zielumwelten angepasst sind, können genetische Distanzen zur Etablierung heterotischer Gruppen benutzt werden. Im subtropisch adaptierten Material wurden zwei heterotische Gruppen, bestehend aus einer Dent- und Flint-Mischpopulation, vorgeschlagen. Bei den tropischen Populationen konnte Population Pop29 in die bereits etablierte heterotische Gruppe A eingeordnet und zwei neue heterotische Gruppen (Pop25, Pop43) vorgeschlagen werden.

Nach den Ergebnissen dieser Studie sind SSR Analysen eine geeignete Methode, um (i) Verwandtschaftsbeziehungen aufzudecken, (ii) den zeitlichen Trend und die vorhandene genetische Diversität in Populationen zu untersuchen und (iii) vielversprechende heterotische Gruppen in Kombination mit Feldversuchen zu etablieren.

10. Acknowledgements

I am very grateful to my academic supervisor Prof. Dr. A.E. Melchinger for his advise, suggestions and support during this thesis work.

Thanks to Prof. Dr. H-.P. Piepho for serving on my graduate committee.

Sincere thanks to Dr. Matthias Frisch for many discussions and his never ending patience in proofreading.

I would like to thank Mrs. B. Boesig and Mrs. S. Meyer for being of great help in organizational matters.

Many thanks to Dr. D. Beck, Dr. M. Bohn, S. Dreisigacker, Dr. M. Frisch, Dr. M. van Ginkel, Dr. D. Hoisington, Prof. Dr. A.E. Melchinger, Dr. S.K. Vasal, Dr. M.L. Warburton, Dr. X.C. Xia, and Dr. P. Zhang for being co-authors of the publications.

I also want to thank Dr. M. Bink, Dr. T.J.L. van Hintum, G. Gort, and Dr. R. Jansen for their support while I was in Wageningen.

Many thanks to S. Dreisigacker, C. Falke, C. Flachenecker, S. Hamrit, Dr. M. Heckenberger, F. Hermann, F. Mauch, H.P. Maurer, V. Merditaj, J. Muminovic, B. Stich, Prof. Dr. H.F. Utz, Dr. P. Zhang, and all unmentioned members within our institute for creating a pleasant work environment.

Curriculum vitae

Name	Jochen Christoph Reif
Birth	09 June 1976 in Rastatt
School education	1982–1986, elementary school (Grundschule Gernsbach).
	1986–1995, high school (Gymnasium Gernsbach). Abitur June 1995.
University education	10/97–09/01, "Agrarbiologie", University of Hohenheim, Stuttgart. Diplom-Agrarbiologe September 2001.
Civil Service	10/95–12/96, parish "San Conrado", Lima, Peru.
Agricultural Experience	04/97–06/97, Südwestdeutsche Saatzucht, Rastatt.
	08/99–10/99, BASF AG, Limburgerhof.
	01/00–12/00, Institute of Animal Breeding, University of Hohenheim, Stuttgart.
	08/00–10/00, CIMMYT, Texcoco, Mexico.
Professional Employment	10/01–03/04, research associate in the Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, Stuttgart. Six month stay at Plant Research International, Wageningen.