Genetic Structure and Diversity Among Maize Inbred Lines as Inferred From DNA Microsatellites

Kejun Liu,* Major Goodman,† Spencer Muse,* J. Stephen Smith,‡ Ed Buckler§ and John Doebley,[1]**

**Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, †Department of Crop Science, North Carolina State University, Raleigh, North Carolina 27695, ‡Crop Genetics Research and Development, DuPont Agriculture and Nutrition, Pioneer Hi-Bred International, Johnston, Iowa 50131, §United States Department of Agriculture-Agricultural Research Service and Department of Genetics, North Carolina State University, Raleigh, North Carolina 27695 and **Laboratory of Genetics, University of Wisconsin, Madison, Wisconsin 53706*

Manuscript received June 10, 2003
Accepted for publication August 20, 2003

ABSTRACT

Two hundred and sixty maize inbred lines, representative of the genetic diversity among essentially all public lines of importance to temperate breeding and many important tropical and subtropical lines, were assayed for polymorphism at 94 microsatellite loci. The 2039 alleles identified served as raw data for estimating genetic structure and diversity. A model-based clustering analysis placed the inbred lines in five clusters that correspond to major breeding groups plus a set of lines showing evidence of mixed origins. A "phylogenetic" tree was constructed to further assess the genetic structure of maize inbreds, showing good agreement with the pedigree information and the cluster analysis. Tropical and subtropical inbreds possess a greater number of alleles and greater gene diversity than their temperate counterparts. The temperate Stiff Stalk lines are on average the most divergent from all other inbred groups. Comparison of diversity in equivalent samples of inbreds and open-pollinated landraces revealed that maize inbreds capture <80% of the alleles in the landraces, suggesting that landraces can provide additional genetic diversity for maize breeding. The contributions of four different segments of the landrace gene pool to each inbred group's gene pool were estimated using a novel likelihood-based model. The estimates are largely consistent with known histories of the inbreds and indicate that tropical highland germplasm is poorly represented in maize inbreds. Core sets of inbreds that capture maximal allelic richness were defined. These or similar core sets can be used for a variety of genetic applications in maize.

MAIZE (*Zea mays* L. ssp. *mays*) inbred lines represent a fundamental resource for studies in genetics and breeding. While maize inbreds are used extensively in hybrid corn production (Anderson and Brown 1952; Troyer 2001), they have also been critical for diverse genetic studies including the development of linkage maps (Burr *et al.* 1988), quantitative trait locus mapping (Edwards *et al.* 1987; Austin *et al.* 2001), molecular evolution (Henry and Damerval 1997; Ching *et al.* 2002), developmental genetics (Poethig 1988; Fowler and Freeling 1996), and physiological genetics (Crosbie *et al.* 1978). Most recently, a set of diverse maize inbreds has been employed to perform the first phenotype-genotype association analyses in a plant species (Thornsberry *et al.* 2001) and to estimate linkage disequilibrium in maize (Remington *et al.* 2001; Tenaillon *et al.* 2001).

The intelligent exploitation of maize inbreds for genetic analyses requires a detailed knowledge of genetic and historical relationships among these lines and an understanding of the partitioning of genetic diversity among them. For example, developmental mutants of maize can exhibit strikingly different phenotypes when assayed in the genetic backgrounds of different maize inbred lines (Poethig 1988). Knowledge of the relationships among lines would help identify a set of inbreds that have maximal diversity for the analysis of the effects of genetic background. Single-nucleotide polymorphism discovery in maize can be optimized by selecting a set of lines that capture the maximum number of alleles or haplotypes. Use of maize inbreds in association analyses requires that population structure among lines be factored into the analysis (Thornsberry *et al.* 2001).

In this article, we analyze the genetic structure and diversity among maize inbred lines using DNA microsatellites or simple sequence repeats (SSRs) and a comprehensive set of 260 inbreds that represent well the diversity available among currently and historically used lines. We show that these lines can be partitioned into five groups, that diversity is greatest among tropical inbreds, that maize inbreds capture ~80% of the allelic diversity in open-pollinated accessions, and that one level of population structure cannot fully explain linkage disequilibrium among inbreds. We also define core sets of inbreds that capture maximal allelic diversity for

[1]*Corresponding author:* Department of Genetics, University of Wisconsin, 445 Henry Mall, Madison, WI 53705. E-mail: jdoebley@wisc.edu

given sample sizes, investigate the relationship between pedigree and genetic distance, and identify the portions of the broader maize germplasm pool from which maize inbreds were derived.

MATERIALS AND METHODS

Plant materials: A set of 260 inbred lines representing a sample of the most important public lines from the United States, Europe, Canada, South Africa, and Thailand, along with lines from the International Center for the Improvement of Maize and Wheat (CIMMYT) and the International Institute of Tropical Agriculture (IITA), was chosen to represent the diversity available among current and historic lines used in breeding. These include essentially all public lines of importance to temperate breeding and many important tropical and subtropical lines. The 260 lines and their pedigrees are listed in supplemental Table S1 at http://www.genetics.org/supplemental/. Seed of most lines are still available from their original sources (see http://www.panzea.org/), but we have also provided seed samples to both the North Central Regional Plant Introduction Station (NCRPIS, Ames, IA) and the National Seed Storage Laboratory (Fort Collins, CO). Most, if not all, lines should be available from the NCRPIS in 2004.

SSR genotyping: The lines were genotyped at Celera AgGen (Davis, CA). The details of the genotyping have been published elsewhere (Romero-Severson *et al.* 2001). Briefly, DNA was extracted from individual plants by the cTAB method, and the microsatellite regions were amplified by PCR with fluorescent-labeled primers. PCR products were size separated on Applied Biosystems (Foster City, CA) fragment analyzers equipped with GeneScan software, and the PCR products were classified to specific alleles (bins) by GeneScan and Genotyper software programs (Romero-Severson *et al.* 2001). We used 100 SSR loci that are evenly distributed throughout the genome. A list of the SSR loci with their chromosomal locations has been deposited as supplemental Table S2 at http://www.genetics.org/supplemental/. Primer sequences are available at the MaizeGDB (http://www.maizegdb.org).

Preanalysis: We began with 264 lines, some of which were assayed two to four times for the 100 SSR loci, giving a total of 339 assays. Of the 33,900 SSR genotypes, 4.3% amplified more than one band per inbred line, perhaps because of residual heterozygosity, contamination, or the amplification of similar sequences in two separate genomic regions. To minimize the effect of contamination, we dropped 7 assays with heterozygosity >0.20, an unexpectedly high value for a maize inbred. Further, 4 other assays, which represented the sole assays for 4 lines, were excluded from the study because their position in a preliminary cluster analysis was strongly discordant with their known pedigrees, suggesting a seed or sample mix-up. We also dropped 4 loci with mean within-line heterozygosity >0.10, suggesting that these loci did not faithfully amplify a single locus or that allele calling was problematic. We dropped 2 loci with availability (defined as 1 − proportion of missing or null data) <0.80, suggesting that the locus could not be amplified in the PCR reaction for many lines. The final data set consists of 260 lines and 94 loci.

We performed multiple SSR assays for some lines. So that each inbred is represented only by a single entry in our data set for statistical analyses, we built consensus genotypes for inbreds that were assayed more than once. The main criterion for constructing the consensus genotype was that any allele with frequency >25% is counted, but if three or more alleles have frequency >25%, then we regard the genotype as missing. The second criterion is that if one assay gave a null phenotype but another gave a visible allele, then the inbred was considered homozygous for the visible allele. Since there was a high degree of concordance among replicate assays, inferred consensus genotypes based on these criteria represent only 1.9% of the final data set.

Summary statistics and tests: We used PowerMarker (Liu 2002) to calculate observed heterozygosity, gene diversity (or expected heterozygosity), number of private alleles, number of group-specific alleles, pairwise *F*-statistics, and stepwise mutation model index. Gene diversity was calculated at each locus as

$$2n(1 - \sum_u p_u^2)/(2n - 1 - f),$$

where p_u is the frequency of the uth allele, n is the sample size, and f is the inbreeding coefficient estimated from genotype frequencies (Weir 1996). Stepwise mutation model index was defined as the maximal proportion of alleles that follow a stepwise mutation pattern (Matsuoka *et al.* 2002a). AMOVA was performed (Excoffier *et al.* 1992).

To evaluate the probability that each of the 260 inbreds would have a unique genotype (fingerprint) for a given number of SSRs, 10,000 random samples of 260 lines were drawn from the empirical distribution of allele frequencies on the basis of the observed data for our 260 inbreds. For these random samples, the probability that all 260 simulated lines had a unique genotype was directly estimated for different numbers of loci. To compare the relationship of pedigree distance and genetic distance, we used a Mantel test (Mantel 1967) by setting the permutation number to 100,000. Pedigree distances were calculated as 1 − Malécot coefficient of coancestry (Malécot 1948), using pedigree information from a variety of sources (see supplemental Table S1 at http://www.genetics.org/supplemental/).

Analysis of genetic structure: Lines were subdivided into genetic clusters using a model-based approach with the software package STRUCTURE (Pritchard *et al.* 2000). Given a value for the number of subpopulations (clusters), this method assigns lines from the entire sample to clusters in a way that Hardy-Weinberg disequilibrium and linkage disequilibrium (LD) were maximally explained. We excluded seven popcorn lines and five sweet corn lines in this analysis (see results). At least six runs of STRUCTURE were done by setting the number of populations (K) from 1 to 10. For each run, burn-in time and replication number were both set to 500,000. The run with the maximum likelihood was used to assign lines to clusters. Lines with membership probabilities ≥0.80 were assigned to clusters; lines with membership probabilities <0.80 for all groups were assigned to a "mixed" group. The three largest clusters were then further subdivided by the same method.

To construct a phylogenetic tree, we used the log-transformed proportion-of-shared-alleles distance that is free of the stepwise assumption, enjoys low variance, and is widely used with multilocus SSR data (Matsuoka *et al.* 2002b). We used the Fitch-Margoliash least-squares algorithm implemented in the computer program Phylip to construct phylogenetic trees (Felsenstein 1993). The tree was rooted using five samples of the maize wild relative, teosinte (*Z. mays* ssp. *parviglumis*) as the outgroup (Matsuoka *et al.* 2002b).

Analysis of allelic richness: We wanted to compare the allelic richness in maize inbreds to that in open-pollinated landrace accessions to estimate the extent to which our set of 260 inbreds captures the diversity present in maize overall. For comparison, we used a previously published data set for 193 maize landrace samples that represent the entire maize germplasm pool (Matsuoka *et al.* 2002b). To compare the allelic richness of inbreds to landraces, we need to adjust for the inbreeding coefficient since inbreds are mostly homozygous

while landraces have a high degree of heterozygosity. We also need to adjust for sample size since our sample has 260 inbreds but only 193 landraces. We used two approaches. First, we compared sets of randomly chosen lines from the inbred and landrace data sets with the same sample sizes. The inbred and landrace genotypes were first broken into alleles to simulate the selfing process. Then, the allele number was counted for randomly drawn samples of size 3–193 in steps of five. Second, we used a parametric simulation to simulate the creation of 260 inbreds from the landraces. The inbreeding coefficient (f) for inbreds was estimated to be 0.965. For each locus, we sampled two alleles with replacement to generate a diploid genotype. If the two alleles are the same, then the simulated inbred is made homozygous. If the two alleles are different, then the simulated inbred is made heterozygous with probability $1 - f$ and made a/a with probability $f/2$ and b/b with probability $f/2$. This procedure was repeated to create 10,000 independent samples of 260 inbreds from which the mean number of alleles and other summary statistics were calculated. The summary statistics for these simulated data were compared with the actual inbred data.

Estimating the historical sources for inbreds: To estimate the historical sources for each inbred group, we used SSR data for 104 representative accessions from four likely historical germplasm pools: Southern Dent, Northern Flint, Tropical Highland maize, and Tropical Lowland maize (supplemental Table S3 at http://www.genetics.org/supplemental/; Matsuoka *et al.* 2002b). We calculated the likelihood of the allelic constitution of an inbred group [*e.g.*, non-Stiff Stalk (NSS)] or specific inbred line, given different proportions of ancestry from the four historical germplasm pools. Assuming that the loci are independent, the likelihood is

$$L(P|n_{lj}, f_{klj}) \propto \prod_{K=1}^{4} \prod_{j=1}^{a_l} \left(\sum_{K=1}^{4} p_k \cdot f_{klj} \right)^{n_{lj}}, \quad 0 < p_k < 1, \sum_{k=1}^{4} p_k = 1,$$

where P is a vector of the proportions of ancestry from the four historical germplasm pools, a_l is number of alleles at the lth locus, f_{klj} is the frequency of the jth allele at the lth locus in the kth population as estimated from the 104 representative landrace lines, n_{lj} is the count of the jth allele at the lth locus for the inbreds group (or line), and p_k is the probability that the allele originated from the kth population. This function was maximized by sequential quadratic programming. Several starting points were chosen to check the global convergence. Standard deviation and confidence interval were inferred from the likelihood surface using established methods (Edwards 1972).

Defining core sets of inbreds: We developed a new algorithm for building core sets of germplasm by maximizing allelic richness using simulated annealing (Kirkpatrick *et al.* 1983). Given the complete set of lines (L), the algorithm works by first randomly selecting a subset of lines (l). Each line has a weight (w) based on the number of private alleles in that line. Next, between 1 and the minimum (l, $L - 1$), additional lines are chosen from the remainder of the complete set (unselected lines) on the basis of their weights and swapped with the same number of the initially selected l lines also chosen on the basis of their weights. The number of alleles (n) is then evaluated and the swap is accepted if it increases n but accepted only with some probability (P) if n is the same or less. The probability of acceptance is dependent on the level of decrease in allelic richness and on the iteration number such that P is larger in earlier iterations. Swapping is continued for a predefined number of iterations. Since P gradually decreases with iterations (time), the method simulates an annealing process. Under this approach, lines with more private alleles have a larger probability to be included in the core set. Our algorithm can also incorporate a weight for the agronomic quality of the inbred and can allow some inbreds to be designated as "conserved" such that they are automatically included in the core set. The details of the algorithm will be given in a separate article (K. Liu and S. Muse, unpublished results).

Linkage disequillibria: The matrix of P values for the pairwise estimates of LD among all 94 SSR loci was evaluated in PowerMarker by the permutation version of Fisher's exact test (Liu 2002). The numbers of locus pairs with LD P values less than threshold values of 0.05, 0.01, 0.001 were counted for the observed data. This analysis was performed on each of the clusters of inbreds defined by the program STRUCTURE as well as the entire set of inbreds. Because the P value of the exact test is affected by sample size and the clusters varied widely in size, we evaluated the effect of sample size on the proportion of significant LD P values by drawing random samples from the entire set of 260 lines of a size equal to the actual number of lines in the cluster. These random samples were drawn without replacement, the exact test was performed on each, and the mean proportion of significant LD P values for 100 replicates was used to compare with the actual results for each cluster.

RESULTS

SSR diversity: We surveyed 260 diverse maize inbred lines using 94 SSRs. The inbreds can be roughly grouped as including 82 tropical lines, 35 temperate Stiff Stalk lines, 131 temperate non-Stiff Stalk lines, seven popcorn lines, and five sweet corn lines. The pedigrees for each line are available online (supplemental Table S1 at http://www.genetics.org/supplemental/). Among the lines, we detected a total of 2039 alleles or an average of 21.7 alleles per locus (Table 1). A large number of private alleles (556 or 27%) are found in only 1 of the 260 inbred lines. Most alleles are at low frequency (Figure 1).

The number of alleles is not equivalent among loci. Loci with dinucleotide repeat motifs have considerably more alleles (average = 23.9) than loci with repeat motifs of three nucleotides or larger (average = 9.9; Table 1). This difference is also seen for gene diversity, with dinucleotide SSRs (average = 0.839) having a higher gene diversity than longer-repeat SSRs (average = 0.707). The mean gene diversity of all SSRs is 0.818.

SSRs are often presumed to follow a stepwise mutation process due to changes in the number of repeats. However, because size differences among alleles are estimated on the basis of the combined molecular weights of the SSR plus its flanking sequences, indels in the flanking sequences can contribute to allelic variation as well (Matsuoka *et al.* 2002a). These indels can cause a violation of the expectation that allele sizes differ strictly by multiples of the repeat motif. We calculated a stepwise mutation index, or the maximal proportion of alleles at a locus that are simple multiples of the repeat motif length. For all 94 loci, the average stepwise mutation model index was 0.832 with dinucleotide SSRs (0.853) showing a higher index than other repeat loci (0.720).

The large number of alleles per locus and the com-

TABLE 1

Summary statistics for all inbreds and each subgroup

Statistics[a]	Overall	TS	Sweet	NSS	Popcorn	SS	Mixed
Sample size	260	58	5	94	7	33	63
Alleles	2039	1268	272	1207	277	535	1321
Alleles per locus	21.7	13.49	2.89	12.84	2.95	5.69	14.05
Type I SSR alleles/locus	23.9	14.71	2.91	13.91	2.97	5.99	15.22
Type II SSR alleles/locus	9.9	7.07	2.80	7.20	2.80	4.13	7.93
Gene diversity	0.82	0.81	0.64	0.78	0.54	0.59	0.82
Type I SSR gene diversity	0.84	0.83	0.64	0.80	0.56	0.61	0.84
Type II SSR gene diversity	0.71	0.68	0.65	0.68	0.45	0.51	0.72
Group-specific alleles	765	305	26	173	16	43	202
Group-specific alleles/line	2.94	5.26	5.20	1.84	2.29	1.30	3.21
Group-specific alleles (%)		24.05	9.56	14.33	5.78	8.04	15.29
Line-specific alleles	556	204	18	121	11	36	166
Line-specific alleles (%)		16.09	6.62	10.02	3.97	6.73	12.57

[a] Type I markers are dinucleotide SSRs and type II markers are SSRs with longer-repeat motifs.

mon occurrence of private alleles suggest that a relatively small number of SSRs would be sufficient to uniquely fingerprint maize inbreds. For the 260 inbred lines that we sampled, the following six loci form a unique profile: *bnlg244*, *bnlg2238*, *bnlg619*, *bnlg1191*, *bnlg1046*, and *dupssr28*. Assuming the allele distribution of our inbred data is representative of all maize inbreds, the probability of sampling 260 independent lines without generating the same genotype for any two lines will be >0.99 by randomly selecting 10 loci. This number is 9 if one uses only dinucleotide SSRs and 12 if one uses longer-repeat SSRs. Thus, very few SSRs are necessary if one wishes to uniquely fingerprint maize inbreds.

Genetic structure of inbred lines: We wished to assess the degree of relatedness among lines and to identify clusters of genetically similar lines. To do this, we used a model-based approach with the program STRUCTURE to subdivide the lines into clusters (Pritchard *et al.* 2000). Five sweet corn lines and seven popcorn lines were assigned to two predefined groups (sweet and popcorn) and were excluded in the STRUCTURE analysis. This was done because a pilot analysis showed that incorporating these 12 lines in the analysis interfered with the ability of STRUCTURE to converge on a robust solution. K (number of populations) = 3 was found to converge well and showed comparable or higher likelihoods than K = 4–10 among runs of the program. We used the run with highest log-likelihood at K = 3 for the observed data to define the model-based groups (Table 2).

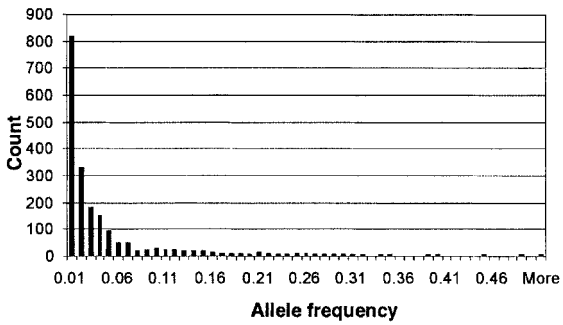

Figure 1.—Histogram of allele frequencies for the 2039 total alleles.

The model-based groups are largely consistent with known pedigrees of the lines (M. M. Goodman and J. S. Smith, personal observation). The largest group has 94 lines, most of which are regarded by breeders as temperate NSS lines. The next group has 58 lines, most of which are either tropical or semitropical (TS) lines. The smallest group has 33 lines, all of which are temperate Stiff Stalk (SS) lines. The remaining 63 lines have <80% membership in any one group and were assigned to a mixed group. Supplemental Table S4 at http://www.genetics.org/supplemental/ shows the proportional membership for these mixed lines in the three groups. Most mixed lines are either NSS-TS or NSS-SS mixtures. Only four lines (Tzi16, Tzi25, Hi27, and CML92) present high membership of TS and SS.

STRUCTURE analysis was repeated to break the three main clusters into subclusters (Table 2). The SS group split into four subgroups, the TS group into five, and the NSS group into seven. Supplemental Table S5 at http://www.genetics.org/supplemental/ shows the proportional membership of the lines in the subgroups for the group to which they belong.

A Fitch-Margoliash "phylogenetic" tree was constructed to further assess the genetic structure of maize inbreds (Figure 2). The tree shows good agreement with the pedigree information and STRUCTURE analysis (see discussion). A version of the tree with the names of the inbreds is available online (supplemental Figure S1 at http://www.genetics.org/supplemental/).

Genetic diversity within inbred groups: Gene diversity

TABLE 2

List of the 260 lines by their model-based groupings

Group	Subgroup[a]	Inbreds
NSS	Hy:T8:Wf9	CI21E, H49, Hy, Mo1W, Pa875, Pa880, T8, Va17, Va14, Va22, Va35, Va102, W64A, Wf9
	M14:Oh43	A619, Gn2, H95, M14, Oh40B, Oh43, Oh43E, PA762, Va26, Va85
	CO109:Mo17	A556, A682, CI.187-2, CO109, CO220, K187, Mo17, MS1334, ND246, W401
	C103	B57, C103, C123, DE2, L317, L1546, NC258, NC262
	Ga:SC	4226, F44, Ga209, GT112, SC357, SC213R, SC213
	NSS-X	38-11, A239, A659, AR4, CM7, CM37, R168, Mo44, MS71, NC260, PA884P, R4, R177, W22
	K64W	33-16, CI.31A, CI.64, CI.66, CI.7, E2558W, Ky21, K55, K64, M162W
	NSS-mixed	A554, A654, B2, B52, B70, B77, B97, B103, CO106, CO125, F6, Fe2, H99, Mt42, N6, Os420, Pa91, R109B, SD44, T234, W153R
SS	B14A	A214N, A632, A634, A635, A665, B14A, B64, B68, CM105, CM174, H91
	B37	B37, B76, H84, NC250
	N28	N28, N28Ht
	B73	A679, A680, B73, B84, B104, B109, NC328, NC372, R229
	SS-mixed	A641, De811, H100, N192, N196, NC294, NC368
TS	TZI	A6, CML52, CML238, CML287, NC358, Q6199, Tzi8, Tzi9, Tzi10, Tzi18
	Suwan	B96, CML69, CML228, CML349, Ki3, Ki9, Ki11, Ki14, Ki44, Ki2007
	CML-late	CML5, CML9, CML61, CML103, CML220, CML254, CML258, CML261, CML264, CML314, Tx601
	CML-early	CML14, CML247, CML311, CML321, CML322, CML331, CML332
	NC	NC296, NC298, NC304, NC336, NC338, NC348, NC350, NC352, NC354
	CML-P	CML10, CML11, CML45, CML277, CML281, CML333, CML341
	TS-mixed	CML38, CML108, NC300, NC356
Sweet corn		Ia2132, Il14H, Il101t, Il677a, P39
Popcorn		HP301, I29, IDS28, IDS69, SA24, Sg18, Sg1533
Mixed		A188, A272, A441-5, A656, B79, B94, B105, B164, C49A, CML77, CML91, CML92, CML218, CML323, CML328, CMV3, CO159, CO255, D940Y, DE3, EP1, F2, F2834T, F7, Hi27, I137TN, I205, IDT, Ki43, Ky226, Ky228, L578, Le23, Le773, M37W, Mo18W, Mo24W, Mp339, MS153, N7A, NC264, NC320, NC360, NC362, NC364, NC366, NC370, Oh7B, Oh603, SC55, SD40, SD46, T232, TEA, Tx303, Tzi11, Tzi16, Tzi25, U267Y, Va99, W117, W117Ht, W182B

The 260 lines in our study are listed with the groups and subgroups from the STRUCTURE analysis. Lines in the mixed group show <80% membership for any group. Seven popcorn lines and five sweet corn lines were assigned into predefined popcorn and sweet corn groups. Within TS, SS, and NSS groups, lines are organized into several distinct subgroups and one mixed subgroup on the basis of analyses with the program STRUCTURE.

[a] The subgroups are named after a defining inbred line(s), principal source (*e.g.*, NC for North Carolina), maturity (early *vs.* late), or mixed for lines that showed <80% membership for any subgroup.

and mean numbers of alleles for the 94 SSRs were calculated for each group of inbreds (Table 1). The TS group is the most diverse with 13.49 alleles per locus and gene diversity of 0.81. NSS has less diversity than TS does, as revealed by the decreased allele number (12.84) and gene diversity (0.78). SS was found to be less diverse than NSS and TS. Our samples of sweet and popcorn include only a few lines, and thus the small numbers of alleles in these groups were expected. In all groups, dinucleotide loci have a much larger allele number than longer-repeat loci. Gene diversity shows a similar trend.

Maize inbreds show a high number of line-specific (556 or 27%) or group-specific (765 or 38%) alleles (Table 1). Far more line- and group-specific alleles are found in the TS group (204 and 305) than in the NSS group (121 and 173) despite a much smaller sample size for TS, indicating far greater diversity in tropical than in temperate inbreds.

An AMOVA revealed that most (90.16%) of the genetic variation resides within groups and only a small percentage resides between groups (8.32%) or within lines (1.51%). Overall F_{st} among groups is 0.086 (95% confidence interval 0.080–0.092) with F_{st} for each locus ranging from 0.02 to 0.17. Pairwise comparisons show a low level of differentiation between TS and NSS (F_{st} = 0.06), but more substantial differentiation between SS and the other groups (Table 3). Popcorn is also highly differentiated from all the other groups. A similar pattern of differentiation among groups is seen using Nei's minimum distance.

Allelic richness of maize inbreds: Comparison of diversity in inbreds to that in open-pollinated landraces shows that the latter possess much greater diversity. For the landraces, the number of alleles (2697 or 28.7 alleles/locus) and overall gene diversity (0.84) are higher than that for the inbreds (2039 or 21.7; 0.82). To compare allelic richness in inbreds *vs.* landraces for equal sample sizes, we randomly selected equal numbers of samples from both germplasm pools (see MATERIALS AND METHODS). This analysis revealed the greater allelic

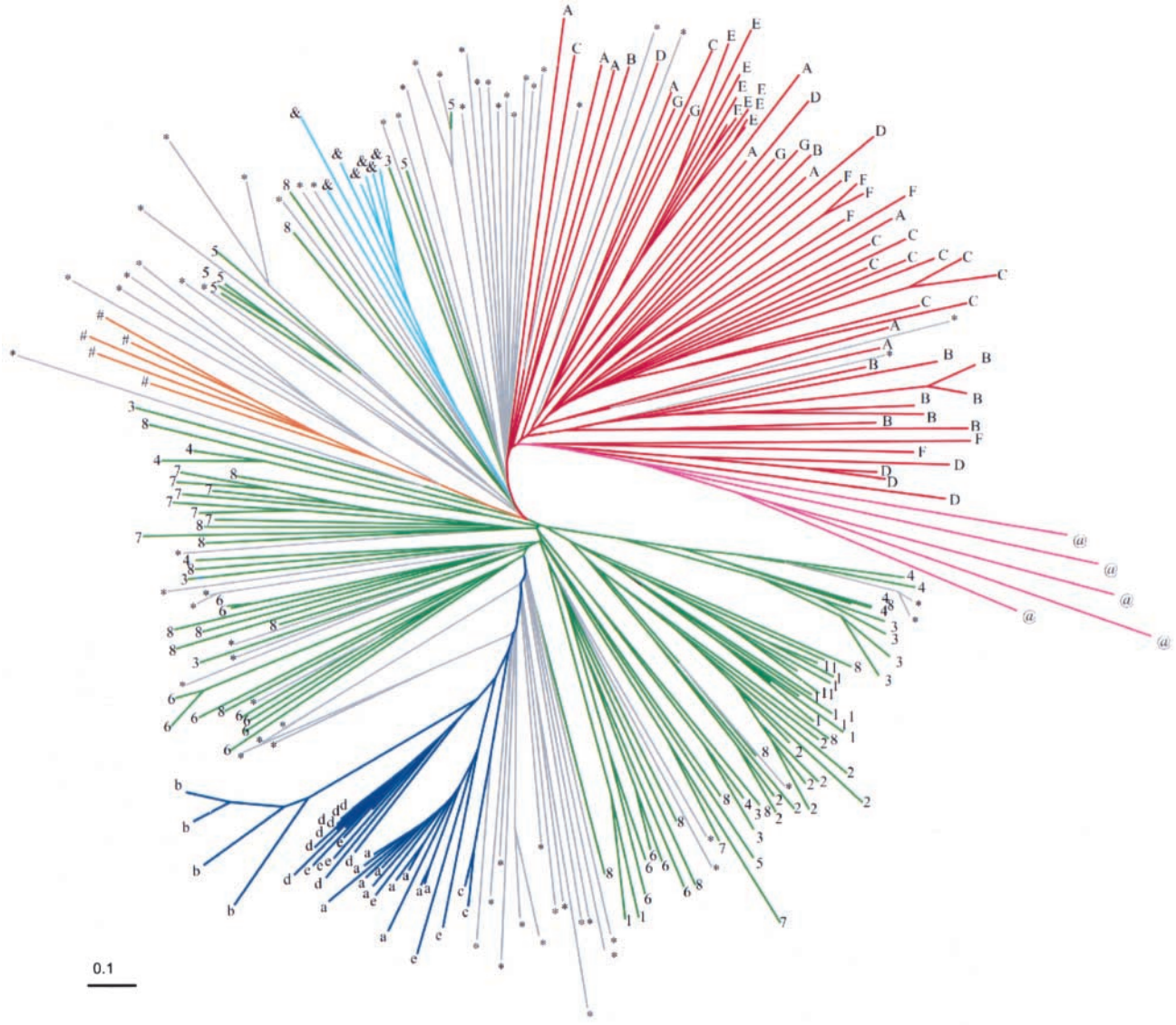

FIGURE 2.—Fitch-Margoliash tree for the 260 inbred lines using the log-transformed proportion of shared alleles distance. The tree was rooted using five teosinte (*Z. mays* ssp. *parviglumis*) samples as outgroups. A version of this tree with the names of the individual inbred lines can be found as supplemental Figure S1 at http://www.genetics.org/supplemental/.

richness in landraces when the samples are equivalent (Figure 3). When the sample size is small (<20), the inbreds capture ~88% as many alleles as the landraces. When the sample size is large (>100), inbreds capture ~78% as many alleles as the landraces.

We also compared allelic richness in inbreds *vs.* landraces, using a parametric simulation. Simulated samples of 260 inbred lines drawn from the landrace gene pool had an average gene diversity of 0.837 (standard error = 0.0015), which is very close to the value for the landrace

TABLE 3

Genetic distances between maize inbred groups

Group	TS	Sweet	NSS	Popcorn	SS
TS	—	0.58	0.29	0.52	0.47
Sweet	0.15	—	0.47	0.62	0.61
NSS	0.06	0.12	—	0.46	0.32
Popcorn	0.15	0.29	0.15	—	0.57
SS	0.18	0.28	0.14	0.31	—

The top diagonal is Nei's minimum distance and the bottom diagonal is pairwise F_{st}.

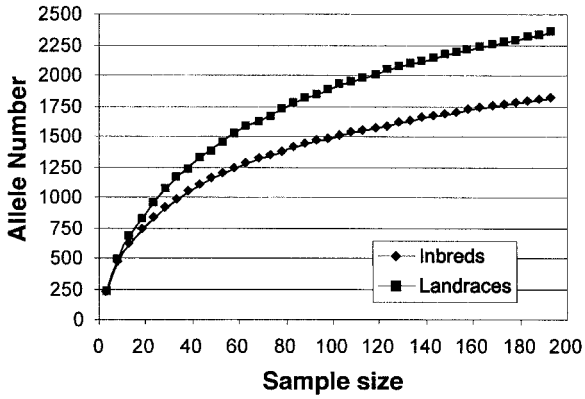

FIGURE 3.—Plots of the expected number of alleles in samples of different sizes. For a given sample size, 1000 replicates were sampled from the inbred or landrace data set without replacement and the genotypes were randomly broken into alleles. Then the mean number was calculated to give the plot from sample size 3 to 193.

sample (0.840). The minimal value of gene diversity in the simulations is 0.832, which is still higher than that of our actual inbred sample (0.820). The mean number of alleles obtained by the simulations is 2292 and the standard error is ~15. The total number for the inbred sample (2039) is not in the 99% confidence interval [2239, 2334], indicating that if one randomly created a set of 260 inbreds from the landrace gene pool, it would contain substantially more allelic diversity than our actual set of 260 inbreds.

Relationship of inbreds to landraces: To understand the relationship between the inbreds and landraces, we estimated the proportion of each inbred group's gene pool that was derived from four different segments of the landrace gene pool (Northern Flint, Southern Dent, Tropical Lowland, and Tropical Highland). TS has its origin mostly from Tropical Lowland (66%) and Tropical Highland (18%; Table 4). NSS and SS show very similar origins, each being composed of roughly equal proportions of Northern Flint, Southern Dent, and Tropical Lowland. Popcorn has a high proportion of Northern Flint germplasm (40%) with most of the rest of its genome coming from Tropical Lowland (26%) and Southern Dent (23%). Sweet corn has the largest contribution from Northern Flint germplasm (72%). Overall, Tropical Highland maize has made a more modest contribution to our set of inbreds than have the other three historical sources. Variances for these estimates are usually small (SD $< 1\%$). Estimates of historical sources for individual inbreds are included in supplemental Table S4 at http://www.genetics.org/supplemental/.

Comparison of SSR and pedigree relationships: A Mantel test shows a highly significant ($P < 10^{-6}$) correlation between pedigree and SSR distance, although the correlation coefficient is relatively small ($r = 0.57$). A plot of pedigree by SSR distances shows a generally strong relationship but with many outliers (Figure 4).

Core sets of inbreds: We defined core sets of inbreds that capture the maximum number of alleles for a given sample size (Table 5). In selecting these sets, we constrained the selection to include 6 lines (A632, B37, B73, C103, Mo17, and Oh43) of high agronomic importance. We also eliminated 8 lines (A654, B2, CM37, CMV3, CO109, I205, Q6199, and R109B) because of poor agronomic quality under our field conditions. Additional core sets of different sizes can be found in supplemental Tables S6 and S7 at http://www.genetics.org/supplemental/. Our study shows that 10 lines capture 28% of the 2039 SSR alleles in the 260 lines, 20 lines capture

TABLE 4

Contributions of historical sources for each model-based inbred group

Group	Tropical Lowland (mean ± SE)	Southern Dent (mean ± SE)	Tropical Highland (mean ± SE)	Northern Flint (mean ± SE)
NSS	0.31 ± 0.01	0.37 ± 0.01	0.05 ± 0.01	0.27 ± 0.01
Popcorn	0.26 ± 0.03	0.23 ± 0.02	0.11 ± 0.03	0.40 ± 0.03
SS	0.32 ± 0.01	0.38 ± 0.01	0.08 ± 0.01	0.23 ± 0.01
Sweet	0.14 ± 0.02	0.06 ± 0.02	0.08 ± 0.02	0.72 ± 0.03
TS	0.66 ± 0.01	0.11 ± 0.01	0.18 ± 0.01	0.04 ± 0.01

The estimates and their standard errors of the historical sources are summarized for each group. These are maximum-likelihood estimates (MLEs). Variance was estimated from the observed Hessian matrix. Because of the MLE's asymptotic normality, the 95% confidence interval can be constructed approximately from the mean ± 1.96 × SE.

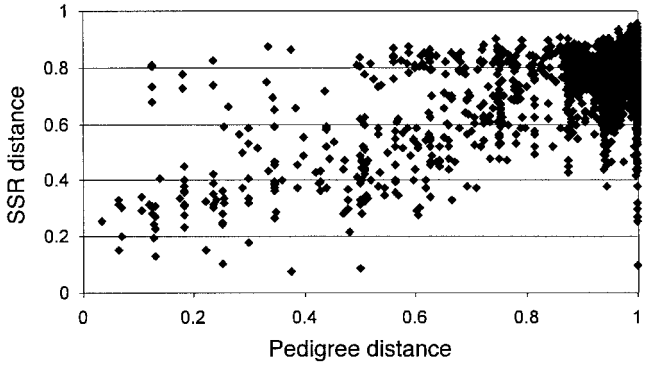

FIGURE 4.—Plot of the SSR proportion of shared allele distance by the pedigree-based (1 − Malécot's coefficient of coancestry) distance for maize inbreds.

46% of the alleles, 30 lines capture 58%, and 50 lines capture 73%. To recover all 2039 alleles, 193 lines were needed. The core sets generally include a large proportion of TS lines as expected since these lines have the greatest allelic richness.

Linkage disequilibria: We assessed extent of LD among SSRs for our sample of inbreds. LD was significant at a 0.01 level between 66% of the SSR marker pairs when all lines were included in the analysis (Table 6). The proportion of significant pairwise LD tests was less within each model-based group. Reduced power to detect LD with fewer lines could contribute to a part of this reduction. However, when we evaluated the percentage of significant pairwise tests in sets of randomly chosen inbreds of the same size as a given group, we observed that sample size alone fails to explain all the reductions (Table 6). This suggests that either linkage or population structure within the NSS, TS, and SS groups contributes to LD. In particular, SS shows a much larger observed LD value, which may be a consequence of the fact that the SS group actually consists of four well-defined subgroups.

TABLE 6

Percentage of SSR locus pairs in LD at a $P = 0.01$ level

Population	No. of lines	Observed % in LD	Expected % in LD[a]
Overall	260	66.05	
NSS	94	19.29	18.91
TS	58	14.48	9.13
SS	33	28.92	4.32

[a] Based on average percentage of all locus pairs showing LD in random samples containing the same number of lines.

DISCUSSION

SSR diversity: Previous studies have shown that maize contains abundant SSRs (SENIOR and HEUN 1993; SENIOR *et al.* 1996, 1998) and that these SSRs are highly polymorphic even among small samples of maize inbreds (CHIN *et al.* 1996; TARAMINO and TINGEY 1996). These pioneering studies were conducted using relatively small numbers of inbreds (9–94) and loci (6–70). We have extended these earlier analyses by using both a large number of SSRs (94) and a much larger number of inbreds (260) that encompass a much greater portion of the maize gene pool. Our analyses uncovered abundant allelic variation with an average of 21.7 alleles per locus over 94 loci. This value greatly exceeds the previously reported values of 5.21 (SENIOR *et al.* 1998), 6.6 (TARAMINO and TINGEY 1996), 4.9 (LU and BERNARDO 2001), and 6.9 (MATSUOKA *et al.* 2002a) alleles per locus. The larger number of alleles observed in the present

TABLE 5

Core sets of inbred lines

Sample size	Alleles obtained	Inbreds
10	579	A632, B37, B73, C103, Mo17, Oh43, CML5, CML52, CML91, Tzi18
20	943	A632, B37, B73, C103, Mo17, Oh43, B96, CML14, CML52, CML91, CML228, CML277, CML281, CML322, Il14H, M37W, Mo18W, Oh7B, Tx601, Tzi8
30	1179	A632, B37, B73, C103, Mo17, Oh43, A272, A441-5, B96, CML5, CML14, CML61, CML77, CML91, CML220, CML228, CML277, CML281, CML311, CML322, CO159, CO255, Il101t, Ky21, M37W, Mo18W, Oh7B, Tx303, Tx601, Tzi8
50	1481	A632, B37, B73, C103, Mo17, Oh43, A272, A441-5, B57, B96, CI.7, CML5, CML14, CML61, CML69, CML77, CML91, CML220, CML228, CML247, CML254, CML261, CML277, CML281, CML311, CML321, CML322, CML328, CML349, CO159, F2, Hi27, I137TN, IDS28, Il14H, K55, Ky21, M37W, Mo18W, NC304, NC348, NC364, Oh7B, Os420, P39, Tx303, Tzi8, Tzi9, Va85, W401

The first six lines (A632, B37, B73, C103, Mo17, and Oh43) were included by default because of their agronomic importance. A654, B2, CM37, CMV3, CO109, I205, Q6199, and R109B were excluded because of poor agronomic performance in our fields in both Raleigh and Florida.

study can be attributed to the larger number of inbreds surveyed, the more diverse selection of inbreds (tropical, subtropical, and temperate), and the inclusion of more dinucleotide repeat SSRs, which tend to be more polymorphic than SSRs with longer-repeat motifs (Vigouroux *et al.* 2002).

We have also observed higher values of gene diversity than those seen in previous analyses of SSR variation in maize inbreds. Gene diversity for our sample of SSRs and inbreds was 0.82 as compared to values of 0.59 (Senior *et al.* 1998), 0.76 (Taramino and Tingey 1996), 0.59 (Smith *et al.* 1997), and 0.62 (Matsuoka *et al.* 2002a). Since estimates of gene diversity are not affected by differences in sample size, the higher value that we observed is likely a function of our use of a greater portion of dinucleotide repeat SSRs and of our more diverse set of inbreds. If one considers only SSRs with trinucleotide or longer-repeat motifs, then gene diversity in our sample (0.71) falls nearer to these previous reports.

We also showed that most maize SSR alleles fit a stepwise mutation model with 83% of the alleles fitting multiples of the length of the repeat motif of their respective loci. The 17% of alleles that deviate from the stepwise pattern likely represent cases where there have been indels in the regions flanking microsatellite repeats (Matsuoka *et al.* 2002a). The failure of these SSRs to fit a stepwise model exactly cautions against the use of models that assume a stepwise mutation process. In particular, estimates of genetic distance such as $(\delta\mu)^2$ (Goldstein *et al.* 1995) or measures of population subdivision based on the stepwise mutation model (Slatkin 1995) would be inappropriate with our data.

Genetic structure: Maize inbreds have a complex history, having been derived from multiple open-pollinated varieties and crosses among the inbreds themselves (Gerdes *et al.* 1993). This history makes it difficult to place maize inbreds into realistic groups that reflect their degree of genetic similarity. Pedigree information provides a useful guide; however, selection and genetic drift during inbreeding can cause considerable discrepancies between pedigree and genetic constitution. Moreover, pedigree information for some inbreds is incomplete, inaccurate, or conflicting.

We used the model-based approach of Pritchard *et al.* (2000) to define natural groups of maize inbreds. In performing this analysis, we discovered that the inclusion of small numbers of sweet (five) and popcorn (seven) lines in the analysis prevented the convergence to a robust solution. Apparently, these two groups are represented by too few lines to form distinct clusters, while at the same time they are too divergent from the other lines to fit into the clusters for those lines. Only when the sweet and popcorn lines were excluded did STRUCTURE converge on a robust solution with three clusters representing the temperate SSs, other temperate lines (NSSs), and TS lines. Thus, along with the predefined sweet and popcorn lines, we classify maize inbreds into five groups. Sixty-three lines did not fit into one of these five groups since they consist of a mixture of two or more of the primary groups. A comparison of genetic distances among the five groups indicates that SS are the most divergent (Table 4), a result consistent with the observation that the SS lines typically provide a strong heterotic response in crosses with other maize inbreds (Hallauer *et al.* 1988).

Inbreds in each of the three model-based groups were analyzed again using STRUCTURE to identify subclusters of related lines (Table 2). The SS group split into four tight subgroups of lines derived from B14, B73, B37, and N28 (see Anonymous 1999). The TS group split into five distinct subgroups with a clear relationship to the origin of these lines. For example, lines in the TZI subgroup are fully tropical and many are from the IITA's streak-breeding-resistance program. The Suwan subgroup consisted mainly of tropical lines that were derived from the Suwan-1 composite population, principally of Caribbean origin (Sriwatanapongse *et al.* 1993), plus B96 from Maíz Amargo of Argentina. The CML-late subgroup is composed of tropical lines tracing back to CIMMYT's late-maturing Tuxpeño composite populations. The CML-early subgroup contained lines derived from CIMMYT's early-maturing (in the tropics) Tuxpeño-related materials and other intermediate-maturity sources. The NC subgroup consists of lines derived from Latin American tropical hybrids. Lines in the subgroup CML-P were largely derived from the La Posta Population 43 (CIMMYT 1987).

The NSS group is organized into seven subgroups that reflect known heterotic groups (Anonymous 1999). The lines in subgroup Hy:T8:Wf9 all trace to these three lines that were important in the era of double-cross hybrids. Lines in subgroup M14:Oh43 all trace to M14, which was an important inbred in the 1940s and 1950s, and Oh43, which is still among our most important breeding sources. Several lines (A556, MS1334, ND246, and W401) with no known relationship to one another were grouped loosely within the CO109:Mo17 subgroup. Mo17 and CO109 represent important, but independent, breeding sources. Subgroup C103 consists mostly of Lancaster germplasm (Hallauer *et al.* 1988) with an additional contribution from B57. Lines within subgroup Ga:SC (Georgia and South Carolina) are mostly southern U.S. germplasm with the notable exception of line 4226. Subgroup NSS-X is a heterogeneous mixture of mostly older lines. The K64W subgroup contains a set of related white lines.

A Fitch-Margoliash tree based on the SSR data shows generally good agreement with the pedigree information and STRUCTURE analysis (Figure 2, supplemental Figure S1 at http://www.genetics.org/supplemental/). There is a general separation of the TS, NSS, and SS lines. Mixed lines are usually located between clusters of TS/NSS/SS lines. Within the SS lines, the four sub-

groups defined by the STRUCTURE analysis are perfectly matched with four clades. For the TS group, the tree has three clades that correspond to subgroups NC, Suwan, and CML-late. For the NSS lines, the tree has three clades that largely correspond to subgroups Hy:T8:Wf9, M14:Oh43, and K64W. All of the sweet corns fall in the same clade, as did all of the popcorns. The European (F2, F7, and EP1) lines and one Canadian (CO255) line are closely grouped together, and this clade is neighbor to the sweet corn clade as expected since all these lines were derived from the Northern Flint landrace of the northern United States and adjacent Canada (Galinat 1971; Doebley *et al.* 1986). NSS-X is also contained within a single large clade, despite the fact that these lines have heterogeneous pedigrees.

Genetic diversity among inbred groups: The amount of genetic diversity within each of the model-based groups is not equivalent. Rather, gene diversity is highest in tropical inbreds (TS), followed by NSS, sweet corn, SS, and popcorn in that order. The greater diversity of the TS lines is again shown by the fact that TS lines contain more alleles than NSS (1268 *vs.* 1207) despite the fact that the sample size for TS was much smaller (58 *vs.* 94). TS lines also possess by far the greatest number of group-specific alleles (305). These data argue strongly that TS inbreds represent an important source of diversity for broadening the genetic base for maize breeding (Goodman 1985; Goodman and Carson 2000).

Of the 2039 alleles, 556 (27%) occur in only one inbred, and 765 alleles (38%) are restricted to a single model-based group of inbreds. These large proportions of private alleles are probably a function of the high mutation rate for maize SSRs (Vigouroux *et al.* 2002), allowing much new allelic variation to arise within lines after their initial development. This feature of maize SSRs contributes to their considerable discriminatory power, enabling one to fingerprint uniquely our entire set of 260 lines with as few as 10 SSRs. This discriminatory power makes SSRs ideal markers for use in varietal identification (Smith *et al.* 1997) and for monitoring gene flow between lines (Dale *et al.* 2002). SSRs can also be used to determine pedigrees in maize inbreds and hybrids but more (*e.g.*, 60 or more SSR loci) are required to trace pedigrees than to provide for unique line identification especially when closely related inbreds are considered (Berry *et al.* 2002).

We also compared diversity in maize inbreds relative to the open-pollinated landraces from which the inbreds were ultimately derived. For this purpose, we used a sample of 193 landrace accessions that represent the entire maize germplasm pool. In particular, we examined the number of alleles captured in our set of 260 inbreds as compared to the number of alleles expected to be captured if these 260 lines represented a random sample of the maize gene pool. The results, whether obtained by a random sampling approach or parametric simulation, revealed a deficit of alleles within the 260 inbreds relative to expectations. For example, a set of 260 inbreds selected at random from the maize gene pool would be expected to capture 2292 alleles on the basis of the parametric simulations while the actual set of 260 lines captures only 2039. This result argues that plant breeders could capture additional diversity by working with landrace accessions (Goodman 1985). It is likely that the landraces contain numerous agronomically useful alleles not represented in the inbreds and advanced populations with which breeders presently work.

Historical sources of maize inbreds: To better understand the relationship between our set of 260 inbreds and the broader maize germplasm pool from which they were derived, we made maximum-likelihood estimates of the portions of four segments of the landrace gene pool (Northern Flint, Southern Dent, Tropical Lowland, and Tropical Highland) represented in the five inbred groups. The results are consistent with historical records, pedigree information, and geography. The temperate NSS and SS are composed of a near-equal mix of Tropical Lowland, Southern Dent, and Northern Flint, although the Northern Flint portion is a bit smaller. Since Southern Dents themselves are thought to have been recently derived from Tropical Lowland germplasm (Galinat 1985; Doebley *et al.* 1988), the high portion of Tropical Lowland germplasm in NSS and SS lines likely represents a tropical contribution that came via the Southern Dents. The observation that NSS and SS are composed of only 25% Northern Flint is consistent with prior observations (Doebley *et al.* 1988).

The sweet corn lines possess the highest portion of Northern Flint, which is consistent with their origin from the flints of the eastern United States (Galinat 1971). The TS lines are composed largely of Tropical Lowland germplasm with a much smaller contribution from Tropical Highland germplasm. Overall, Tropical Highland germplasm is the least well represented among our 260 maize inbreds, although it is probably the most diverse segment of the maize gene pool (Matsuoka *et al.* 2002b; J. Doebley, unpublished data). This result suggests that maize breeders should consider maize of the Tropical Highlands as a source of new allelic diversity. However, Highland Tropical germplasm has consistently proven difficult to integrate into temperate breeding programs because of its susceptibility to heat stress and fungal pathogens when grown at lower elevations (Goodman 1999).

In addition to our estimates of historical contributions to the inbred groups, we have estimated the historical sources for each of our individual 260 inbreds (supplemental Table S4 at http://www.genetics.org/supplemental/). The only inbred in our sample with a high proportion of Tropical Highland germplasm (72%) is CML349, which is a Tropical Highland inbred line. This again points to the possibility of using Tropical Highland germplasm to increase diversity within maize inbreds. The top four lines in terms of Northern Flint contribu-

tion (IA2132, IL14H, IL101t, and P39) are all sweet corn. Some European lines (F2, F7) and one Canadian line (CO255) also have >50% Northern Flint origin. Va35, a southern U.S. line, was found to have the largest Southern Dent proportion (63%).

Pedigree *vs.* genetic distance: Previous studies using molecular markers have generally shown a strong correlation between molecular marker and pedigree-based distance measures (Smith and Smith 1992; Bernardo *et al.* 1997; Smith *et al.* 1997; Bernardo *et al.* 2000; Bernardo and Kahler 2001). Nonetheless, estimates of relatedness based on pedigree data can differ from those based on molecular marker data (Bernardo and Kahler 2001). Calculations of relatedness based upon pedigree data are dependent upon the assumptions that both parents contribute an equal number of alleles (*i.e.*, no selection, mutation, or genetic drift) and that the pedigree data are accurate. Another assumption is that founder genotypes (genotypes for which no further pedigree information on ancestors is available) are unrelated by pedigree. All of these assumptions can be violated.

We observed a highly significant correlation between pedigree- and SSR-based distances, although a much weaker correlation (0.57) than that seen in some previous studies. For example, Smith *et al.* (1997) reported a correlation of 0.81 between SSR and pedigree distances for maize inbreds. Since their sample of inbreds included many commercial lines with detailed pedigrees, it is not surprising that they observed a stronger correlation than we did with our more diverse set of lines. Similarly, Bernardo *et al.* (2000) observed a correlation of 0.92 between SSR and pedigree distances using a small set of public inbreds with well-documented pedigrees. Our study also differs from these two prior studies in using a higher proportion of dinucleotide SSRs, which with their higher mutation rate could weaken the correlation between SSR and pedigree distance.

Linkage disequilibrium: Overall, 66% of SSR pairs exhibited significant LD. Smaller percentages of SSR pairs showed significant LD within the model-based groups, due in part to reduced statistical power with the smaller sample sizes. Sets of inbreds chosen at random from the full set of 260, but of the same size as one of the model-based groups, showed less LD than the actual model-based groups themselves (Table 6). This result suggests that the observed LD is largely due to population structure (or linkage) within groups as opposed to higher-level population structure among the entire set of lines. In particular, we observe a large excess of SSR pairs in LD within SS (29%) as compared to the expected number (4%), suggesting either considerable population structure or linkage effects among SS lines. This result is in apparent disagreement with those of Remington *et al.* (2001) who found that higher-level structure makes an important contribution to LD and who observed the least evidence for LD within SS lines. This disagreement is likely a result of different sampling. Remington *et al.* (2001) had fewer inbreds that were chosen in a manner to avoid closely related lines, while our larger set included many closely related lines, especially closely related SS lines.

Perspective: There is a heightened awareness of the necessity for maintaining genetic diversity for the study of natural variation and for crop improvement. However, when stocks are placed in germplasm banks without an adequate understanding of the amount and distribution of genetic variation within those stocks, potential users of these resources are confronted with the difficulty of choosing a diverse and representative selection from long lists of essentially anonymous accessions. In this article, we have shown that maize inbreds possess a great depth of allelic diversity. This diversity is not distributed randomly among the lines, but rather diversity is structured into five groups along breeding group (SS *vs.* NSS) and ecological (temperate *vs.* tropical) axes. Similarly, the amount of diversity is not equivalent among groups, but rather tropical-subtropical inbreds possess greater diversity than their temperate counterparts. It is also clear that allelic diversity in some portions of the broader maize gene pool is not well represented in available inbreds. In particular, we found that the diversity in Tropical Highland maize is poorly represented among available inbreds, suggesting that Tropical Highland germplasm could be tapped to identify new alleles of agronomic importance. Finally, to aid researchers working with maize, we have defined both core sets of maize inbreds and a method for choosing core sets to best represent diversity among a set of inbreds. These results should help maize researchers to make more informed choices of inbreds for research and breeding.

We thank Bruce Weir for comments on the manuscript. This work was supported by the U.S. National Science Foundation grant DBI-0096033.

LITERATURE CITED

Anderson, E., and W. L. Brown, 1952 Origin of Corn Belt maize and its genetic significance, pp. 124–148 in *Heterosis—A Record of Researches Directed Toward Explaining and Utilizing the Vigor of Hybrids*, edited by J. W. Gowen. Iowa State College Press, Ames, IA.

Anonymous, 1999 *MBS Genetics Handbook*. Mike Brayton Seeds, Ames, IA.

Austin, D. F., M. Lee and L. R. Veldboom, 2001 Genetic mapping in maize with hybrid progeny across testers and generations: plant height and flowering. Theor. Appl. Genet. **102:** 163–176.

Bernardo, R., 2001 Breeding potential of intra- and interheterotic group crosses in maize. Crop Sci. **41:** 68–71.

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

Bernardo, R., A. Murigneux, J. P. Maisonneuve, C. Johnsson and Z. Karaman, 1997 RFLP-based estimates of parental contribution to F_2- and BC_1-derived maize inbreds. Theor. Appl. Genet. **94:** 652–656.

Bernardo, R., J. Romero-Severson, J. Ziegle, J. Hauser, L. Joe *et al.*, 2000 Parental contribution and coefficient of coancestry

among maize inbreds: pedigree, RFLP, and SSR data. Theor. Appl. Genet. **100:** 552–556.

Berry, D. A., J. D. Seltzer, C. Xie, D. L. Wright and J. S. C. Smith, 2002 Assessing probability of ancestry using simple sequence repeat profiles: applications to maize hybrids and inbreds. Genetics **161:** 813–824.

Burr, B., F. A. Burr, K. H. Thompson, M. C. Albertson and C. W. Stuber, 1988 Gene mapping with recombinant inbreds in maize. Genetics **118:** 519–526.

Chin, E. C., M. L. Senior, H. Shu and J. S. Smith, 1996 Maize simple repetitive DNA sequences: abundance and allele variation. Genome **39:** 866–873.

Ching, A., K. S. Caldwell, M. Jung, M. Dolan, O. S. Smith *et al.*, 2002 SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet. **3** (19): 1–14.

CIMMYT, 1987 *CIMMYT Report on Maize Improvement 1982–83*, pp. 1–78. Centro Internacional de Mejoramiento de Maíz y Trigo, Chapingo, Mexico.

Crosbie, T. M., J. J. Mock and R. Pearce, 1978 Inheritance of photosynthesis in a diallel among eight maize inbred lines from Iowa Stiff Stalk Synthetic. Euphytica **27:** 657–664.

Dale, P. J., B. Clarke and E. M. G. Fontes, 2002 Potential for the environmental impact of transgenic crops. Nat. Biotech. **20:** 567–574.

Doebley, J., M. M. Goodman and C. W. Stuber, 1986 Exceptional genetic divergence of the Northern Flint corns. Am. J. Bot. **72:** 64–69.

Doebley, J., J. F. Wendel, J. S. C. Smith, C. W. Stuber and M. M. Goodman, 1988 The origin of cornbelt maize: the isozyme evidence. Econ. Bot. **42:** 120–131.

Edwards, A. W. F., 1972 *Likelihood: An Account of the Statistical Concept of Likelihood and Its Application to Scientific Inference*. Cambridge University Press, Cambridge, UK.

Edwards, M. D., C. W. Stuber and J. F. Wendel, 1987 Molecular-marker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. Genetics **116:** 113–125.

Excoffier, L., P. Smouse and J. Quattro, 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics **131:** 479–491.

Felsenstein, J., 1993 *PHYLIP—Phylogeny Inference Package*, Version 3.5c. Department of Genetics, University of Washington, Seattle.

Fowler, J. E., and M. Freeling, 1996 Genetic analysis of mutations that alter cell fates in maize leaves: dominant *Liguleless* mutations. Dev. Genet. **18:** 198–222.

Galinat, W. C., 1971 The evolution of sweet corn. Univ. MA Amherst Coll. Agric. Agric. Exp. Stn. Res. Bull. **591:** 1–20.

Galinat, W. C., 1985 Domestication and diffusion of maize, pp. 245–282 in *Prehistoric Food Production in North America*, edited by R. I. Ford. University of Michigan, Ann Arbor, MI.

Gerdes, J. T., C. F. Behr, J. G. Coors and W. F. Tracy, 1993 *Compilation of North American Maize Breeding Germplasm*. Crop Science Society of America, Madison, WI.

Goldstein, D. B., A. Ruiz Linares, L. L. Cavalli-Sforza and M. W. Feldman, 1995 Genetic absolute dating based on microsatellites and the origin of modern humans. Proc. Natl. Acad. Sci. USA **92:** 6723–6727.

Goodman, M. M., 1985 Exotic maize germplasm: status, prospects, and remedies. Iowa State J. Res. **59:** 497–529.

Goodman, M. M., 1999 Broadening the genetic diversity in breeding by use of exotic germplasm, pp. 139–148 in *Genetics and Exploitation of Heterosis in Crops*, edited by J. G. Coors and S. Pandey. Crop Science Society of America, Madison, WI.

Goodman, M. M., and M. L. Carson, 2000 Reality vs. myth: corn breeding, exotics and genetic engineering. Proc. Annu. Corn Sorghum Res. Conf. **55:** 149–172.

Hallauer, A. R., W. A. Russell and K. Lamkey, 1988 Corn breeding, pp. 463–564 in *Corn and Corn Improvement*, edited by G. F. Sprague and J. W. Dudley. Crop Science Society of America, Madison, WI.

Henry, A., and C. Damerval, 1997 High rates of polymorphism and recombination at the *Opaque-2* locus in cultivated maize. Mol. Gen. Genet. **256:** 147–157.

Kirkpatrick, S., C. D. Gelatt and M. P. Vecchi, 1983 Optimization by simulated annealing. Science **220:** 671–680.

Liu, J., 2002 *Powermarker—A Powerful Software for Marker Data Analysis*. North Carolina State University Bioinformatics Research Center, Raleigh, NC (www.powermarker.net).

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

Malécot, G., 1948 *Les Mathématiques de l'Hérédité*. Masson & Cie, Paris.

Mantel, N., 1967 The detection of disease clustering and a generalized regression approach. Cancer Res. **27:** 209–220.

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

Matsuoka, Y., Y. Vigouroux, M. M. Goodman, G. J. Sanchez, E. Buckler *et al.*, 2002b A single domestication for maize shown by multilocus microsatellite genotyping. Proc. Natl. Acad. Sci. USA **99:** 6080–6084.

Poethig, R. S., 1988 Heterochronic mutations affecting shoot development in maize. Genetics **119:** 959–973.

Pritchard, J. K., M. Stephens and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics **155:** 945–959.

Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt *et al.*, 2001 Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA **98:** 11479–11484.

Romero-Severson, J., J. S. C. Smith, J. Ziegle, J. L. Hauser and G. Hookstra, 2001 Pedigree analysis and haplotype sharing within diverse groups of *Zea mays* L. inbreds. Theor. Appl. Genet. **103:** 567–574.

Senior, M. L., and M. Heun, 1993 Mapping maize microsatellites and polymerase chain reaction confirmation of the targeted repeats using a CT primer. Genome **36:** 884–889.

Senior, M. L., E. C. L. Chin, M. Lee, J. S. C. Smith and C. W. Stuber, 1996 Simple sequence repeat markers developed from maize sequences found in the GENBANK database: map construction. Crop Sci. **36:** 1676–1683.

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.

Slatkin, M., 1995 A measure of population subdivision based on microsatellite allele frequencies. Genetics **139:** 457–462.

Smith, J. S. C., E. C. L. Chin, H. Shu, O. S. Smith, S. J. Wall *et al.*, 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.

Smith, O. S., and J. S. C. Smith, 1992 Measurement of genetic diversity among maize hybrids—a comparison of isozymic, RFLP, pedigree, and heterosis data. Maydica **37:** 53–60.

Sriwatanapongse, S., S. Jinahyon and S. Vasal, 1993 *Suwan-1: Maize From Thailand to the World*. Centro Internacional de Mejoramiento de Maíz y Trigo, Chapingo, Mexico.

Taramino, G., and S. Tingey, 1996 Simple sequence repeats for germplasm analysis and mapping in maize. Genome **39:** 277–287.

Tenaillon, M. I., M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley *et al.*, 2001 Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc. Natl. Acad. Sci. USA **98:** 9161–9166.

Thornsberry, J. M., M. M. Goodman, J. Doebley, S. Kresovich, D. Nielsen *et al.*, 2001 *Dwarf8* polymorphisms associate with variation in flowering time. Nat. Genet. **28:** 286–289.

Troyer, A. F., 2001 Temperate corn, pp. 393–466 in *Specialty Corns*, edited by A. Hallauer. CRC Press, Boca Raton, FL.

Vigouroux, Y., J. S. Jaqueth, Y. Matsuoka, O. S. Smith, W. D. Beavis *et al.*, 2002 Rate and pattern of mutation at microsatellite loci in maize. Mol. Biol. Evol. **19:** 1251–1260.

Weir, B. S., 1996 *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.

Communicating editor: J. A. Birchler