American Journal of Botany 95(10): 1240–1253. 2008.

Population structure and genetic diversity of New World maize races assessed by DNA microsatellites[1]

Yves Vigouroux,[2,3] Jeffrey C. Glaubitz,[2] Yoshihiro Matsuoka,[4] Major M. Goodman,[5] Jesús Sánchez G.,[6] and John Doebley[2,7]

[2]Department of Genetics, University of Wisconsin, Madison, Wisconsin 53706 USA; [3]Institut de Recherche pour le Développement, UMR141, Montpellier 34394 France; [4]Fukui Prefectural University, Matsuoka-Cho, Yoshida-gun, Fukui, 910-1195, Japan; [5]Department of Crop Science, North Carolina State University, Raleigh, North Carolina 27695 USA; and [6]Centro Universitario de Ciencias Biológicas y Agropecuarias, Universidad de Guadalajara, Zapopan, Jalisco, Mexico CP45110

Because of the economic importance of maize and its scientific importance as a model system for studies of domestication, its evolutionary history is of general interest. We analyzed the population genetic structure of maize races by genotyping 964 individual plants, representing almost the entire set of ~350 races native to the Americas, with 96 microsatellites. Using Bayesian clustering, we detected four main clusters consisting of highland Mexican, northern United States (US), tropical lowland, and Andean races. Phylogenetic analysis indicated that the southwestern US was an intermediary stepping stone between Mexico and the northern US. Furthermore, southeastern US races appear to be of mixed northern flint and tropical lowland ancestry, while lowland middle South American races are of mixed Andean and tropical lowland ancestry. Several cases of post-Columbian movement of races were detected, most notably from the US to South America. Of the four main clusters, the highest genetic diversity occurs in highland Mexican races, while diversity is lowest in the Andes and northern US. Isolation by distance appears to be the main factor underlying the historical diversification of maize. We identify highland Mexico and the Andes as potential sources of genetic diversity underrepresented among elite lines used in maize breeding programs.

Keys words: diversification; domestication; genetic diversity; microsatellites; races; *Zea mays* subsp. *mays.*

Between five and ten thousand years ago, humans domesticated virtually all major crop species currently used in modern agriculture. During this process, cultivated plants underwent domestication bottlenecks (Tanksley and McCouch, 1997) that generally reduced their gene diversity relative to their wild ancestors. After domestication, cultivated plants spread from their center of origin with the spread of agriculture and agricultural societies. For example, maize (*Zea mays* L. subsp. *mays*) was first domesticated in southern Mexico from teosinte (*Z. mays* L. subsp. *parviglumis*) between 6000–10 000 years ago (Matsuoka et al., 2002), but then spread to North America and South America. During this expansion, maize became adapted to diverse climates and soil types such as the deserts of the southwestern United States and the high elevations of the Andes mountains. Plant morphology also diversified greatly during this process. For example, ear shape varies from grenade shaped in much of the Andean maize to long and slender in races from the northern United States.

During the spread of maize cultivation, different maize lineages acquired distinct genetic and morphological characteristics. Sets of plants sharing particular characteristics have been classified into races. Members of a race have pronounced similarities not only in morphological phenotype and geographical distribution, but also in genetic, cytological, physiological, and agronomic characteristics (Wellhausen et al., 1952; McClintock et al., 1981). A race has been locally sown year after year, and its diversity has been dynamically maintained by local farmers. With the occurrence of modern agriculture, new varieties have been introduced, often originating from crosses among elite inbred lines. These new varieties as a group usually contain less genetic diversity but also have higher yields; thus, they have replaced old local races in some places (Sánchez G. et al., 2000a). As a result of this encroachment, diversity that was previously dynamically maintained by traditional farming methods could be lost, possibly impairing the future development and improvement of the domesticated plant.

A global analysis of maize race genetic diversity is a first step to understand its organization and how its geographical distribution has been shaped during maize domestication and dissemination. Initial work on the genetic relationships between races was based upon chromosomal knob morphology (McClintock et al., 1981). However, knobs are associated with meiotic drive (Buckler et al., 1999) and thus may have undergone convergent selection. Isozyme markers were employed in subsequent studies, usually focusing on a particular country or area (e.g., Doebley et al., 1985, 1986; Bretting et al., 1990; Sánchez G. et al., 2000a, 2006). A global analysis of isozyme diversity at 23 genetic loci in 1080 accessions from more than 300 races of maize from the Americas was conducted by Sánchez G. et al. (2000b). In this study, focusing on genetic diversity rather than on phylogenetic relationships and dispersal history, allelic diversity was concentrated in Mesoamerica, and expected heterozygosity was lowest within northern flint and Andean races.

In addition to studies with isozymes, several studies of maize races have been carried out at the DNA level. In their microsatellite-based study of the origin of maize, Matsuoka et al. (2002) examined a broad sample of 193 race accessions covering its entire pre-Columbian distribution and developed a scenario for the spread of maize through the Americas. Microsatellite markers have also been used in a fine-scale study of the genetic structure of maize races in the Valley of Oaxaca,

[1] Manuscript received 18 March 2008; revision accepted 22 July 2008.

This work was supported by U. S. National Science Foundation grants DBI-0096033 and DBI-0321467. The authors thank J. Liu and F. Chevenet for assistance with data analysis.

[7] Author for correspondence (e-mail: jdoebley@wisc.edu), phone: 608-265-5803, fax: 608-262-2976

doi:10.3732/ajb.0800097

Mexico (Pressoir and Berthaud, 2004a) and in a study of Mexican races (Reif et al., 2006). Both microsatellite (Lia et al., 2007) and DNA sequence (Freitas et al., 2003; Jaenicke-Deprés et al., 2003) data have been employed in studies of maize archeological specimens. Finally, microsatellite allele frequencies were analyzed in bulk DNA samples from each of 144 maize race accessions from throughout the Americas, with the primary purposes of controlling for genetic structure in an association analysis of the *dwarf8* gene (Camus-Kulandaivelu et al., 2006) and elucidating the origin of European maize races (Dubreuil et al., 2006).

However, a comprehensive analysis of the maize races of the Americas at the DNA level is still missing. To provide such an analysis, we have supplemented the maize race portion of the data set of Matsuoka et al. (2002) by adding genotypes for 752 additional accessions for a common set of 96 microsatellite loci. This new, combined data set, consisting of genotypes for 964 plants from 945 accessions, covers almost all known races of maize in the Americas. Our aim is to provide a complete picture of the organization of maize race genetic diversity in the Americas to further elucidate the historical, pre-Columbian, spread of maize throughout this region. In addition, we seek to characterize the extent and patterns of more recent, post-Columbian, translocations of race germplasm.

MATERIALS AND METHODS

Plant materials and SSR loci—To the maize race portion of the data set of Matsuoka et al. (2002), we added 771 additional maize plants from 752 additional accessions. Hence, the full maize race data set analyzed here encompassed 964 individual maize plants from 945 different accessions, chosen to cover the pre-Columbian range of maize. These 945 different accessions correspond to 310 different races, race cultivars, or tribes. The sampled accessions encompass almost all of the approximately 350 described races in the Americas, originating from geographic locations ranging from the south of Chile to Canada and from the Andean mountains to the Caribbean islands (Fig. 1). However, corn belt dents were not included in this study because they are more recent in origin. The 310 races were each represented by between one and 14 sampled accessions. In turn, 926 of the 945 accessions were represented by DNA from a single plant, with the remaining 19 accessions represented by two plants. In addition, five *Z. mays* subsp. *parviglumis* plants from the Balsas River Basin, where maize domestication is thought to have occurred (Matsuoka et al., 2002), were included to serve as an outgroup in the phylogenetic analyses (described later). Passport data for all plants, with descriptions of the main collections that were the ultimate sources of the studied germplasm, are available in Appendix S1 (see Supplemental Data with the online version of this article).

We used 96 microsatellite or simple sequence repeat (SSR) loci in this study, distributed throughout the genome (http://www.maizegdb.org/ssr.php). Genotyping was performed at Celera AgGen (Davis, California, USA). The genotyping methods were as previously published (Matsuoka et al., 2002). We used the same set of SSRs as in Matsuoka et al. (2002), except for three markers (bnlg1288, bnlg1931 and phi096) that were excluded because they had >15% missing data in the full data set. The list of SSRs employed, their repeat type and maize chromosomal bin location are provided in Appendix S2 (see Supplemental Data with the online version of this article). The primer sequences are available at the website MaizeGDB (http://www.maizegdb.org/ssr.php). The full data set analyzed here is available online from the website at http://www.panzea.org/data_sets.html. The overall rate of missing data in this data set is 5.2%.

Analysis of genetic relatedness between accessions—*Model-based clustering*—For the analysis of population structure and detection of hybrids, we used the program Structure 2.1, which implements a Bayesian, model-based clustering algorithm (Pritchard et al., 2000). In this analysis, genotyped individuals are allocated to a predetermined number of populations (K) in a manner that minimizes Hardy–Weinberg and linkage disequilibrium within each population; individuals are also allowed to be products of admixture between two or more of the populations. The method employs a Markov chain Monte Carlo algorithm to estimate the allele frequencies in each of the K populations and, for each individual, the proportion of its genome derived from each population (q_K). We assumed that all loci were independent and in linkage equilibrium, and we did not use any prior information about the origin of each plant. We routinely employed 1 000 000 iterations for estimation after a burn-in period of 30 000 iterations. We set Alphapropsd, the standard deviation of the distribution from which proposals for the admixture proportion are drawn, at 0.50 to allow better mixing. At least five independent runs were assessed each time for each fixed number of clusters greater than one. For this analysis, we employed only loci with less than 10% missing data (84 loci). The run with the maximum likelihood was used to assign plants to clusters. We also calculated ΔK for each value of K (except for the maximum K tested), according to Evanno et al. (2005). We attributed a plant to a given cluster when the proportion of its genome in the cluster (q_K) was higher than an arbitrary cutoff value of 80%, the same threshold employed in previous studies (Liu et al., 2003; Fukunaga et al., 2005). After the first analysis including all plants, additional Structure analyses were performed to detect possible substructure within each of the main clusters detected. In this case, because of the narrower genetic base within a cluster, allele frequencies were allowed to be correlated among the subclusters. Although differentiation between subclusters within a cluster as quantified by F_{ST} was low (between 2.9 and 5.3%, see Results), the large number of loci employed

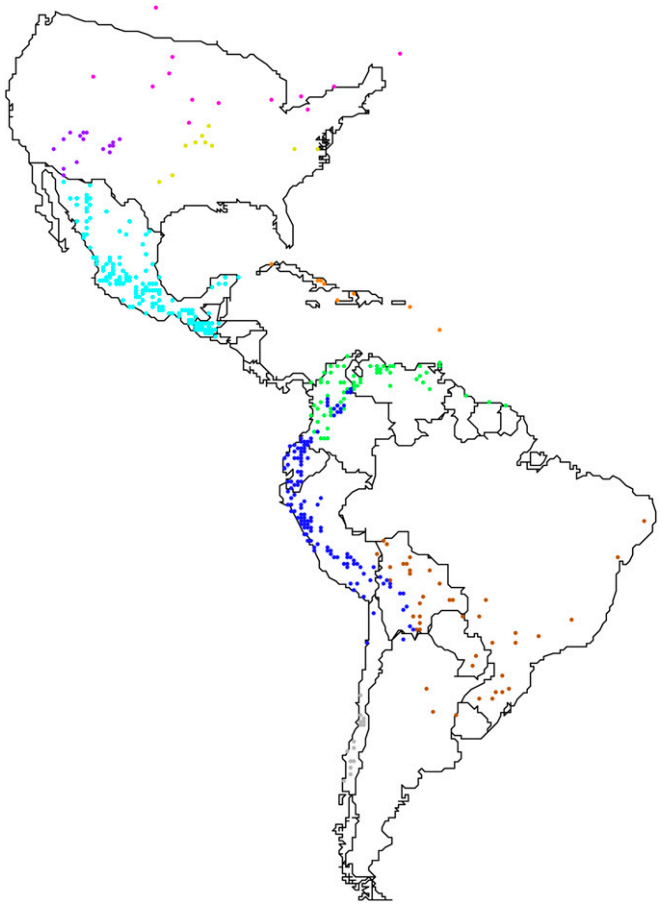

Fig. 1. Geographical distribution of the comprehensive set of maize (*Zea mays* subsp. *mays*) race accessions analyzed in this study. The color scheme matches that of the phylogeny in Fig. 4A. The distinction between Andean accessions (blue) in Columbia from northern South American accessions (green) in that country was based upon elevation (cutoff: 1750 m a.s.l.). Similarly, a 1100 m a.s.l. cutoff distinguished Andean accessions in western Bolivia and western Argentina from those of lowland Middle South America (brown). Forty-seven accessions (representing 48 of the 964 maize plants sampled) lacked latitudinal and longitudinal coordinates and are not shown on the map.

in this analysis (84 loci) should allow the detection of subtly differing subclusters, if they do exist.

Phylogenetic analysis—A phylogenetic tree of individual plants was constructed based on the natural log transformation of the proportion of shared alleles distance (lnPSAD), calculated using the program PowerMarker (Liu and Muse, 2005). The phylogenetic tree was constructed using the neighbor joining algorithm implemented in the computer program PHYLIP (Felsenstein, 2005); this algorithm was used because of its high computational speed, a necessity in light of the large number of plants studied. A second phylogenetic tree was also constructed to show the relationships among the subclusters of plants identified by Structure, based upon a matrix of Nei's genetic distance (Nei, 1972) between subclusters, calculated by PHYLIP. This second dendrogram was constructed using the FITCH algorithm in PHYLIP, which provides a more accurate representation of the distance matrix at the cost of slower computational speed relative to neighbor joining (Felsenstein, 2005); because there were a limited number of subclusters, computational speed was not a limitation in this case. Nei's genetic distance was employed instead of lnPSAD for this second dendrogram because it is a more appropriate measure of the divergence of populations (as opposed to the dissimilarity of individuals). Bootstrap support for this second tree was determined by resampling loci 1000 times. Both dendrograms were formatted and colored with the aid of the program TreeDyn (Chevenet et al., 2006).

Map construction—To illustrate the Structure results, we created geographical maps using the programs ArcView and ArcView Spatial Analyst (ESRI). Separate maps were constructed for each main cluster identified by Structure. Spatial analyses were performed on the basis of a grid consisting of 50 × 50 km cells; this chosen cell size was a compromise between fineness of resolution and computational burden. First, for each cell containing one or more sampled plants, the q_K estimates of the plants were averaged. The cell average was then projected onto a circle having the cell midpoint as its center and a radius of 500 km. This radius was chosen because it filled in the main maize growing areas nicely, without the resulting maps appearing too coarse. Final projected cell values were then recalculated as averages of all of the circles within which each cell midpoint fell.

Genetic diversity analysis—The genetic diversity parameters number of alleles (i.e., allelic richness), frequency of the most frequent allele, and gene diversity were compared among the main clusters identified by Structure, using only those plants that were assigned to each cluster. To compare the number of alleles and the frequency of the most frequent allele among groups (clusters) of different sizes, we used a rarefaction method similar to that of Petit et al. (1998). For clusters other than the smallest cluster, the average number of alleles (or the frequency of the most frequent allele) for each locus in each cluster was calculated as the mean of a thousand resamplings at the smallest cluster sample size (i.e., the number of plants assigned to the smallest cluster). Because gene diversity (i.e., expected heterozygosity) is far less sensitive to sample size (Petit et al., 1998), rarefaction was not used for this statistic. Diversity measures (number of alleles and frequency of the most frequent allele after rarefaction and "raw" gene diversity) were then compared between each pair of groups using a Wilcoxon signed-rank test, with diversity measures for the two groups paired across loci; this test was carried out with the software SYSTAT (Cranes Software International, San Jose, California, USA). Analysis of molecular variance (AMOVA; Excoffier et al., 1992; Excoffier, 2007) was performed using the program Arlequin (Excoffier et al., 2005); statistical significance of each variance component was assessed based upon 10c000 permutations of the data.

Mantel tests—Mantel tests (Smouse et al., 1986) were performed comparing the matrix of genetic distance (lnPSAD) between individual plants with three other matrices. The first was a matrix of geographical distances calculated from latitude and longitude of each sampled accession (where available). The second was a matrix of differences of altitude between accessions. Finally, a matrix based upon race names was constructed by assigning a value of zero if two plants had the exact same race name or one otherwise. All possible pairwise combinations of matrices not involving the genetic distance matrix were also tested. For each pair of matrices, the statistical significance of their correlation was determined from 1000 permutations of the rows and columns of one of the matrices.

Core sets—To find the optimal, "core set" of individuals of a given sample size that capture a maximal number of alleles, we used a line selection algorithm based upon simulated annealing that is described in Liu et al. (2003) and implemented in PowerMarker (Liu and Muse, 2005). The analysis was performed for sample sizes varying from 10 to 300, with one run per sample size. We also determined the minimum number of plants needed to capture 50% and 80% of the total number of sampled alleles. Twenty independent runs were performed to assess the minimum sample size and to determine which accessions should be included to reach the 50% and 80% targets.

RESULTS

Structure analysis—In the Structure analysis for all 964 sampled plants, the number of clusters (K) was varied from one to six, with 10 replicate runs performed for all K values except $K = 1$, where only one run was performed (because only one outcome is possible in this case). The highest likelihood was obtained when K was set to four. Beyond $K = 4$, the likelihood decreased slightly and then stayed constant (Fig. 2). Hence, there was no difficulty determining the K value with the maximum likelihood ($K = 4$). However, using the method of Evanno et al. (2005), maximal ΔK occurred at $K = 2$ (Fig. 2), with the next largest peak at $K = 4$. At $K = 2$, maize races were divided into Andean vs. non-Andean races. Because the large Andean group is one of the most divergent (see phlyogenetic results) and we are also interested in the natural groupings of maize races from outside of the Andes, we selected the K value with the maximal likelihood ($K = 4$) over that with maximal ΔK ($K = 2$). Furthermore, the large number of loci employed in this analysis, leading to firm rejection of the null hypothesis of no structure, seems to cause ΔK to be artifactually maximal at $K = 2$ (discussed more later).

With the arbitrary cutoff value of 80% ancestry for assignment, 544 plants (56%) were attributed to one of the four clusters: 235 plants to an Andean cluster, 187 plants to a tropical lowland cluster, 87 plants to a highland Mexican cluster and 35 plants to a northern United States (US) cluster. The average altitudes of the accessions assigned to the four clusters are 2200 m, 570 m, 1850 m, and 330 m a.s.l., respectively. A geographical representation of the Structure results (Fig. 3) highlights the geographical basis of the four detected clusters.

A large number of plants (420) appeared to have ancestry from more than one cluster, having q_K values of less than 80% for all four clusters. These mixed-ancestry plants included 35 of the 38 plants from southwestern US accessions (races from Arizona and New Mexico). The vast majority (32) of the 38 southwestern US plants appeared to have roughly equal mixtures of highland Mexican and northern US ancestry (q_K between 25% and 75% for both clusters), with very little or no ancestry in the other two clusters ($q_K < 15\%$). Furthermore, varieties of the southeastern United States (undersampled in this study) tended to have mixed northern US (northern flint) and tropical lowland ancestry (q_K between 30% and 70% for both clusters). Regions of apparent mixed-ancestry are clearly visible on the map in Fig. 3: the mixed highland Mexican and northern US ancestry of the southwestern US races can be seen by comparing Fig. 3A and 3D, while the mixed tropical lowland and northern US ancestry of the southeastern US races can be seen by comparing Fig. 3B and 3D. Moreover, plotting the Structure results on these maps allowed us to confirm that germplasm from some races has been translocated: there is evidence of northern US ancestry in some races of Chile and Argentina (Fig. 3D). Accessions from the Chilean and Argentinean races Dulce Golden Bantam, Dulce Evergreen, Cristalino Norteño, and Cateto Sulino Precoce were assigned to the North American cluster. Translocation of northern US germplasm to Chile was previously documented (Timothy et al., 1961).

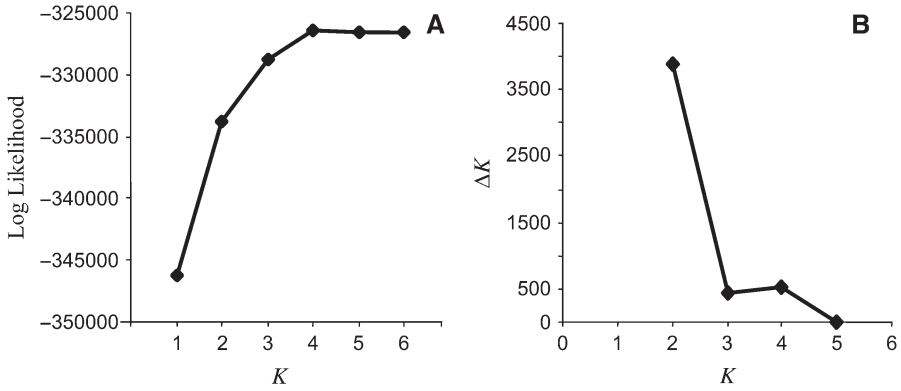

Fig. 2. Plots of (A) the log likelihood, and (B) ΔK, from the Structure analysis of the full set of sampled plants. For the log likelihood plots, only the maximum log likelihood from among the ten replicate runs performed at each K is plotted (except for $K = 1$, where only one run was performed).

Substructure analysis—The results from the substructure analysis, allocating races within each of the four major clusters into subclusters, are provided in Appendix S3 (see Supplemental Data with the online version of this article). Notably, the ΔK method of Evanno et al. (2005) *always* favored $K = 2$, just as in the main structure analysis in which all sample plants were included (described earlier). This was in contrast to the maximal likelihood, which had a distinct peak in each analysis for a specific K greater than two. Maximal ΔK at $K = 2$ hence appears to be an artifact resulting from markedly low likelihoods for $K = 1$ in all cases. The large number of loci we employed in these analyses (84 loci) seems to have provided tremendous power to reject the null hypothesis of no structure within either the entire set of studied plants or within each main cluster. Hence the likelihoods for $K = 1$ were always extremely low relative to those for other values of K. Because the Evanno method focuses exclusively on the change in slope, this caused ΔK to be artificially maximal at $K = 2$ in all cases. The simulations upon which the Evanno et al. (2005) method is based employed only 5 or 10 SSR loci and thus had much lower power to reject $K = 1$. Hence, they likely do not apply to this study.

Phylogenetic analysis—The phylogenetic analysis based upon individual plants (Fig. 4A) resulted in a cogent scenario for maize race evolution. Races from Mexico are basal, being most closely related to the outgroup, *Z. mays* subsp. *parviglumis*. Furthermore, the tree clearly indicates that southwestern US races (from Arizona and New Mexico) were derived from Mexican races and that northern US races were in turn derived from the southwestern US races. Also, the dendrogram indicates that maize expanded into northern South America from Central America, then into the Caribbean from northern South America, most likely via Trinidad and Tobago. Most accessions from the Yucatan Peninsula grouped in the Mexican clade, rather than in the northern South American or Caribbean clades (results not shown); hence, the possibility that maize was first brought to the Caribbean from Yucatan does not seem to be supported by the tree. Races of the Andes appear to have been derived from those

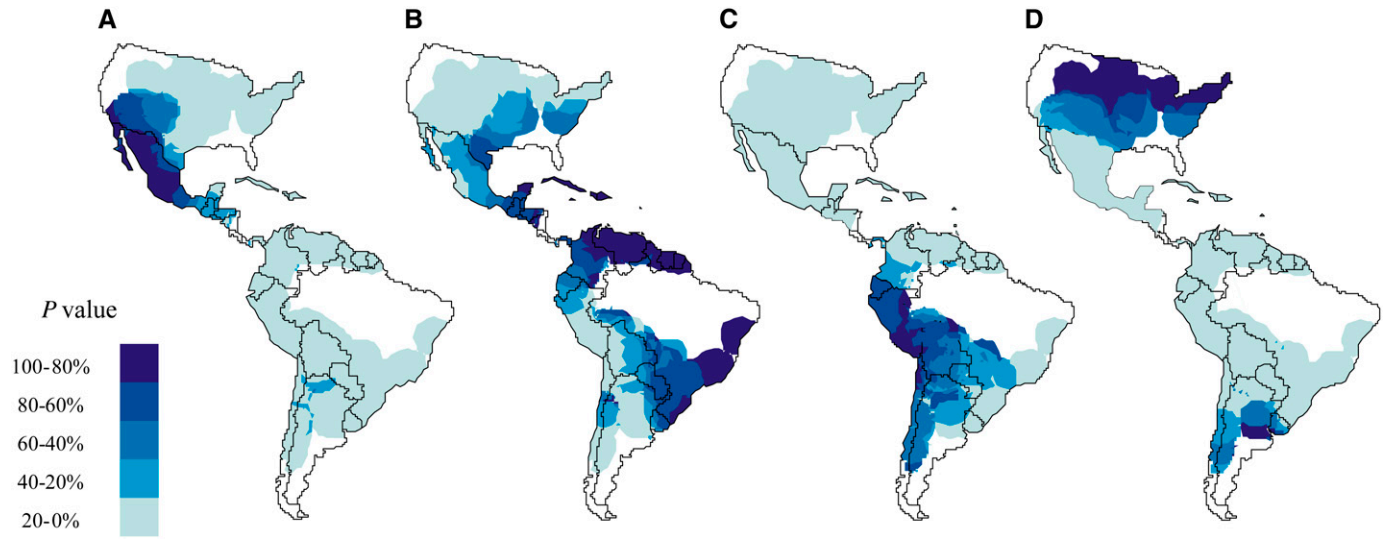

Fig. 3. Geographical representation of maize race ancestral composition estimated by the program Structure. A map is shown for each of the four main clusters detected, on which with the percentage of ancestry in that cluster (q) is plotted: (A) highland Mexican, (B) tropical lowland, (C) Andean, and (D) northern US.

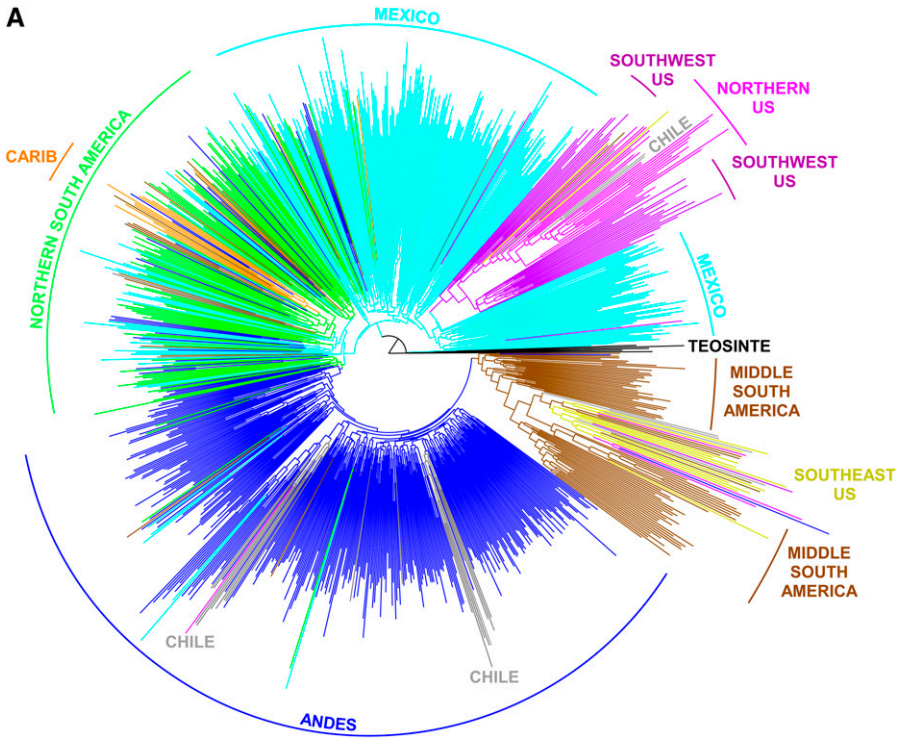

Fig. 4. Phylogenies of maize races. (A) Individual plant neighbor joining dendrogram based upon the log-transformed proportion of shared alleles between individuals, colored according to the geographical origin of each accession, using the color scheme in Fig. 1. (B) The same dendrogram as in A, but recolored according to the four main race clusters identified by the program Structure, with individuals of mixed ancestry in gray. (C) FITCH dendrogram showing the evolutionary relationships among the 18 subclusters identified by Structure, based upon Nei's (1972) genetic distance between subclusters. Races from the southwestern United States were also included as an additional taxon. Clades with greater than 85% bootstrap support are indicated. Both dendrograms were rooted with a teosinte outgroup consisting of five *Zea mays* subsp. *parviglumis* plants (black).

of northern South America, and races of lowland middle South America (Bolivia, Argentina, Paraguay, and Uruguay), in turn, appear to have been derived from those of the Andes.

Several post-Columbian translocations of maize races are apparent on the individual-based tree (Fig. 4A). The races of central Chile, though geographically isolated (Fig. 1), appear from the individual-based tree to have diverse origins: two separate Chilean clades descended from Andes material and, as found in the Structure analysis, most of the remaining Chilean accessions appear to have descended from material translocated from the northern United States. These long distance translocations most likely occurred after European colonization. The two accessions of the Columbian race Maiz Dulce both grouped with the Caribbean material, and hence possibly represent a translocation from the Caribbean to the Andes region. Similarly, other Andean races that grouped near the Caribbean material in the northern South American clade often include Cubano in their name, which verifies that they do indeed represent translocations. Cuban maize is thought to have been spread widely by agricultural extension agents before World War II (D. H. Timothy, North Carolina State University, personal communication). Some Mexican Yucatan and Middle South American accessions also grouped with the northern South American/Caribbean clade; it is probable that once all of the maize race regions were established, germplasm exchange began to occur at zones of secondary contact (e.g., between Middle South America and northern South American, and between the Yucatan Peninsula and the Caribbean).

It would appear from the individual-based tree that races of the southeastern United States are predominantly middle South American in ancestry; however, this is most likely an artifact resulting from post-Columbian translocation of southeastern US germplasm to Brazil and its subsequent introgression into Brazilian material (discussed later). Two southeastern US accessions are better classified as northern flint, and three northern US accessions appear to be better classified as southeastern US; beyond this, however, northern flint ancestry in races of the southeastern US is not apparent from the tree (in contrast to the Structure results—discussed later).

Although an initial, a priori attempt was made to subdivide races from Mexico and Guatemala into seven separate geographical and altitudinal subgroups (consisting of Guatemalan highlands, Guatemalan lowlands, central Mexican highlands, northern Mexican highlands, lowland northwest Mexico, lowland northeast Mexico, and the Yucatan Peninsula), these subgroups did not resolve into separate clades in this dendrogram. Hence, Mexican and Guatemalan races are colored the same (cyan) in Fig. 4A, and this individual-based tree was not informative with regard to the precise geographical location within Mexico where the initial domestication and diversification of maize took place. However, this question has already been addressed by Matsuoka et al. (2002), using a much larger sample of teosinte accessions (and a more focused sample of maize races). Although Dubreuil et al. (2006) were able to resolve two groups of Mexican races with SSRs—the Mexican pyramidal and northern groups previously resolved with isozymes (Sánchez G. et al., 2000a)—their SSR results indicated that they are weakly distinguished at the DNA level.

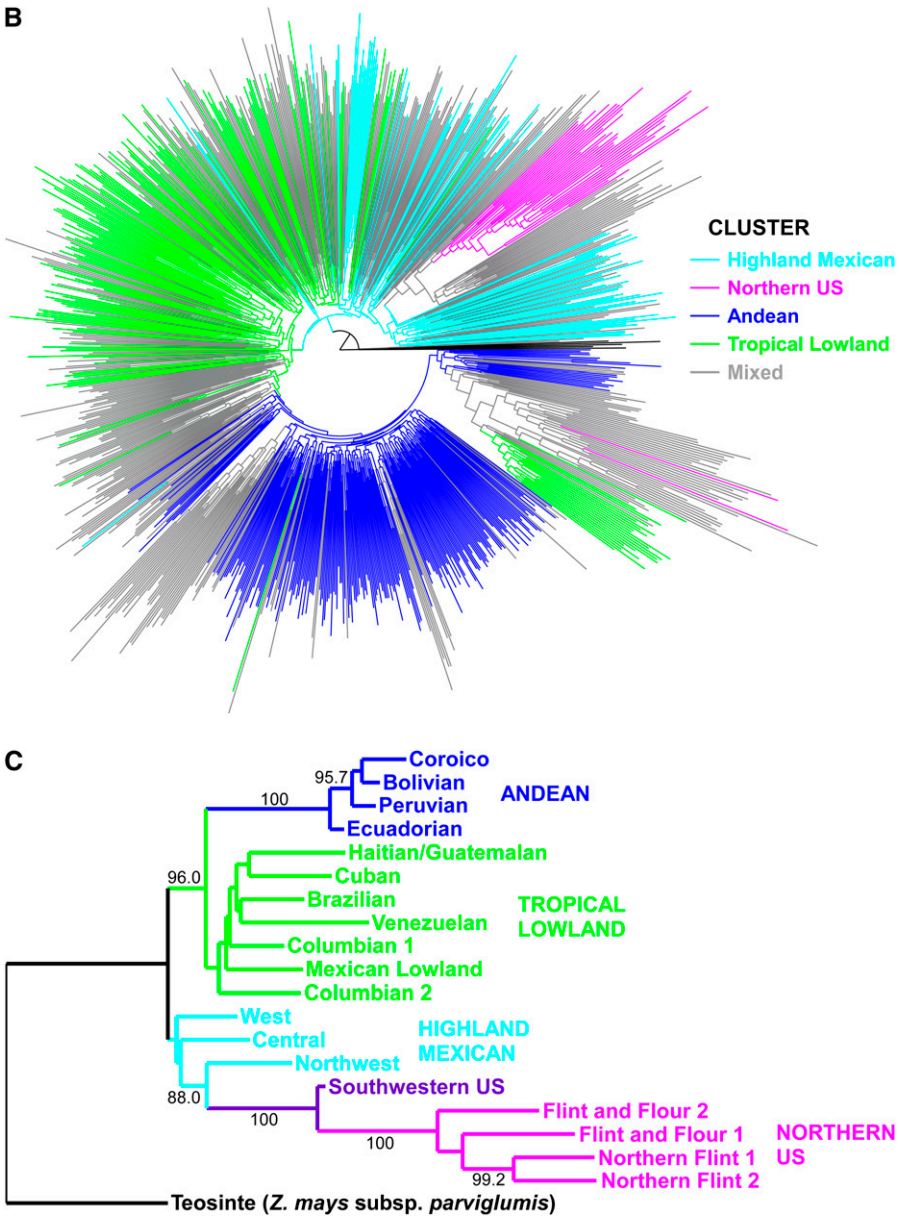

Fig. 4. Continued.

Comparison of phylogenetic and Structure results—To facilitate comparison of the results of the individual-based phylogenetic analysis with those of the Structure analysis, the individual-based tree presented in Fig. 4A was recolored to reflect the four main clusters found by Structure, with individuals of apparent mixed ancestry being colored gray (Fig. 4B). The two analyses agreed in most respects and provided complementary information where they differed. The main difference was that some of the Middle South American accessions, brown in Fig. 4A, were classified with the tropical lowland cluster by Structure and hence are green in Fig. 4B. All these middle South American accessions were placed in the Brazilian subcluster of the tropical lowlands by Structure: this particular green clade in Fig. 4B is composed of 21 of the 24 accessions in the Brazilian subcluster in the substructure analysis (Table S3.1 in online Appendix S3 in Supplemental Data). The remainder of the accessions in the middle South American clade in Fig. 4A were classified by Structure as having mixed Andean and tropical lowland ancestry (and hence are gray in Fig. 4B). Together, the phylogenetic and Structure results indicate that the middle South American region is a zone of secondary contact between Andes material from the west and tropical lowland/northern South America material that arrived via the Brazilian coast. Phylogenetic analysis alone is ill suited to uncovering the reticulate pattern of evolution expected to result from this type of introgression pattern, because in a bifurcating dendrogram, the admixed material will often appear to have been derived from one or the other source population, but not both. In such cases, Structure analysis is complementary.

Complementary information was also provided by the Structure results and the individual-based phylogenetic analysis in regard to the ancestry of races of the southeastern U.S and subsequent translocation of some of that material. Although most accessions of the southeastern US races (except for two) were placed within the middle South American clade in the individual-based phylogeny, the Structure results indicated that these same accessions are comprised of mixed tropical lowland and northern flint ancestry. When the Structure results and the phylogenetic results are considered together along with historical information on the history of southeastern US and Brazilian maize races, the following scenario emerges. Southeastern US races were initially formed as a mixture of northern flint and tropical lowland genetic material, with the tropical lowland material most likely coming from lowland Mexico or the Caribbean (Brown and Anderson, 1948; Doebley et al., 1988; Goodman and Brown, 1988). Subsequently, after the Civil War, southeastern US maize races were brought to Brazil by immigrants from the southern US (Brieger et al., 1958; Paterniani and Goodman, 1978; Sánchez G. et al., 2007), and were then introgressed into preexisting Brazilian germplasm, thus explaining the seemingly anomalous affinity of southern US races with Brazilian and middle South American material in the individual-based dendrogram (Fig. 4A). The limitations of bifurcating dendrograms for describing hybridization and admixture (reticulate evolution) and our very limited sample of southeastern US accessions probably contributed to this apparent anomaly. Hence, it would be interesting to revisit the history of southeastern US races and their relationship with northern flint, tropical lowland, and Brazilian germplasm in a future study with a more comprehensive sampling of southeastern US accessions (to the extent that existing germplasm collections allow). Notably, middle South American accessions that fell *within* the southeastern US clade in the individual-based dendrogram (brown taxa within the predominantly yellow clade in Fig. 4A) belonged to the races Canario de Ocho (from Argentina), Dente Branco Riograndense (from Brazil) and Hickory King (from Brazil), all known to be closely related to US races (Sánchez G. et al., 2007). In addition, three accessions of Dentado Comercial from Chile and a single accession of Cristalino Grande from Chile all fell within the southeastern US clade (gray taxa within the predominantly yellow clade in Fig. 4A), indicating that not only northern flint material, but also southeastern US germplasm was translocated to Chile.

Although some accessions from Mexico and Guatemala (from races such as Bofo, Celaya, Comiteco, Conejo, Dzit Bacal, Jala, Mixeño, Nal-Tel, Negro de Tierra Caliente, Olotón, Ratón, Serrano Mixe, Tabloncillo, Tehua, Tepecintle, Tuxpeño, Zamorano Amarillo, and Zapalote Grande) were placed inside the northern South American clade in the phylogenetic analysis (cyan-colored taxa within the predominantly green clade in Fig. 4A), the Structure results suggested that additional low elevation accessions from Mexico and Guatemala (from races such as Conejo, Dzit Bacal, Motozinteco, Nal-Tel, Negro de Tierra Caliente, Olotillo, Olotón, Tepecintle, Tuxpeño, Vandeño and Zapalote Chico) also belong in the tropical lowland cluster (green taxa in Fig. 4B that, in Fig. 4A, are cyan and fall within the predominantly cyan-colored Mexican and Guatemalan clade). Beyond a certain degree of stochasticity inherent in both methods, we have no explanation for this particular discrepancy. Neither technique deals well with hybridization, and both treat poorly represented taxa less well than widely represented types. However, their relative concordance on a wide range of maize accessions, most of which are represented here by single-plant samples is remarkable. Furthermore, the two analyses do fully concur with respect to placement of the southwestern US races. The intermediate phylogenetic position of the southwestern races between the Mexican and northern flint races is concordant with their classification by Structure as admixed highland Mexican and northern flint.

To further place the Structure results into an evolutionary context, we constructed a FITCH tree based on Nei's genetic distance between the subclusters discerned by Structure (Fig. 4C; the results of the subclustering analysis upon which the dendrogram in Fig. 4C is based are presented in online supplementary Appendix S3). Accessions from the southwestern US were included as an additional taxon on this second dendrogram to further test if races from the northern US directly descended from those in the southwestern US. This FITCH tree clearly indicates that the southwestern races are intermediary between the highland Mexican and northern US races, with 100% bootstrap support of the grouping of the southwestern US races with the northern US races. Furthermore, the subclusters grouped together into main clusters in exactly the manner expected from the Structure results, with strong bootstrap support in most cases. The scenario in which the tropical lowland races descended from the lowland races of Mexico, while races of the Andes in turn descended from those of the tropical lowlands was again supported. Although the dendrogram suggests that tropical lowland maize and highland Mexican maize descended from a common ancestor, this ancestral maize is clearly more similar to highland Mexican maize than tropical lowland maize. Likewise, Andean maize appears to have descended from an ancestral population that is quite similar to current tropical lowland maize. Founder effects upon the formation of the southwestern US, northern US, and Andes groups are evident as long branches on this tree and are associated with somewhat reduced genetic diversity (discussed later).

Diversity analysis between clusters—Based on the 96 loci, the average gene diversity for the entire sample of 964 plants was 0.83. In total, 3752 alleles, or an average of 39 alleles per locus, were detected. To understand how diversity during maize diversification has evolved, we compared gene diversity and allelic richness between the four different main clusters identified in the Structure analysis (Table 1), using the Wilcoxon signed-rank test.

Gene diversity and number of alleles—The gene diversity of the highland Mexican cluster was slightly higher than that of the tropical lowland cluster (Table 1), but the difference was not significant ($P = 0.31$). In contrast, the gene diversity of the Andean cluster was lower than that of both the tropical lowland ($P < 0.001$) and highland Mexican ($P < 0.001$) clusters. Similarly, the gene diversity of the northern US cluster was lower than that of both the highland Mexican ($P < 0.001$) and tropical lowland ($P < 0.001$) clusters but not significantly different from that of the Andean cluster ($P = 0.28$). A similar trend was observed for the average number of alleles per locus (after rarefaction). Allelic richness was highest in the highland Mexican cluster followed by the tropical lowland cluster, but was not significantly different between these two ($P = 0.17$). Allelic richness of the Andean cluster was lower that that of both the highland Mexican ($P<0.001$) and tropical lowland ($P < 0.001$) clusters, but higher than that of the northern US cluster ($P = 0.005$).

Likewise, the northern US cluster had less allelic diversity than both the highland Mexican ($P < 0.001$) and tropical lowland ($P < 0.001$) clusters.

Frequency of the most frequent allele—A larger proportion of loci had high frequency (>0.5) alleles in the Andean cluster than in either the highland Mexican ($P < 0.001$) or the tropical lowland ($P < 0.001$) clusters (Table 1, Fig. 5). No difference was detected between the Andean and the northern US clusters ($P = 0.78$) nor between the tropical lowland and highland Mexican clusters ($P = 0.41$). The northern US cluster had a larger number of loci with high frequency alleles than either the highland Mexican cluster ($P < 0.001$) or the tropical lowland cluster ($P < 0.001$).

AMOVA—The analysis of molecular variance (Table 2) indicated that a low percentage of variation was partitioned either among races or among clusters: only 7.6 and 7.5%, respectively. Only 3.2% of the variation was attributed to differences among subclusters within clusters. A high level of variation was observed among plants within either races or subclusters (24 and 28%, respectively), but most of the variation was found within plants (62–68%). The high diversity within clusters was associated with large and significant F_{IS} inside each cluster: F_{IS} was 0.33, 0.33, 0.29, and 0.44 for the highland Mexican, tropical lowland, Andean, and northern US clusters, respectively (Table 1). However, partitioning each cluster into the subclusters uncovered by the Structure analysis accounted for only small proportions of this diversity: F_{ST} after partitioning each cluster into subclusters was 0.029, 0.039, 0.027, and 0.053 for the highland Mexican, tropical lowland, Andean, and northern US clusters, respectively (Table 1). Differentiation among the four clusters as measured by pairwise F_{ST} was also relatively modest: 0.037 between the highland Mexican and Andean clusters, 0.063 between the Andean and tropical lowland clusters, and 0.025 between the highland Mexican and tropical lowland clusters. Differentiation was higher between the northern US cluster and the three other clusters (Andeans, highland Mexican, and tropical lowland) with F_{ST} estimates of 0.148, 0.116 and 0.175, respectively.

Mantel tests—The relationships uncovered by the pairwise Mantel tests are summarized in Fig. 6. A low but significant correlation was observed between the genetic distance and race name matrices ($R = 0.065$, $P < 0.001$). As expected, race names, originally defined by eco-geographical and cultural criteria, were correlated both with altitude ($R = 0.165$, $P < 0.001$) and with geographical distance ($R = 0.068$, $P < 0.001$). Genetic distance was also correlated both with altitude ($R = 0.044$, $P < 0.001$) and with geographical distance ($R = 0.41$, $P < 0.001$). The latter correlation between geographical and genetic distance was by far the strongest of those tested. To further characterize this correlation, we calculated the average genetic distance for all pairs of plants falling within each 100-km distance class (Fig. 7A). Average genetic distance increased steadily as a function of geographical distance and then began to plateau at around 4000 km. This isolation by distance pattern was well described by the polynomial function: $y = -5 \times 10^{-9}x^2 + 8.5 \times 10^{-5}x + 1.4$, where x is the geographic distance class and y is the average genetic distance ($R^2 = 0.98$). A finer scale analysis was also performed for distances up to 300 km, using a shorter distance class interval of 10 km (Fig. 7B). Here, genetic distance rapidly increased between zero and 50 km and then increased more slowly. This finer scale pattern of isolation by distance was well described by the logarithmic function: $y = 0.032 \cdot \ln x + 1.28$, again where x is the geographic distance class and y is the average genetic distance ($R^2 = 0.71$).

Core sets—In the total sample of 964 plants, 3752 different alleles were observed. For sample sizes varying from 10 to 300, we identified, for each sample size, the combination of plants that would capture the maximum number of alleles from our total sample (Fig. 8). This range of sample sizes was chosen to give an overall sense of the shape of the curve, but ignoring extremely small (<10) or extremely large sample sizes (>300). The smallest sample set that accounted for 50% of the total number of alleles consisted of only 41 accessions (Table 3). So with only 4.2% of the total sample of plants, we could capture 50% of the total number of alleles sampled. The ancestral contribution of the different clusters to this sample set was 34%, 32%, 18% and 16% from the highland Mexican, tropical lowland, Andean, and northern US clusters, respectively. To obtain a sample representing 80% of the total number of 3752 alleles, we needed a larger sample of 172 accessions (Table 3). The ancestral contribution to this larger sample was 35%, 34%, 21% and 10% from the four clusters, respectively.

DISCUSSION

This study describes genetic diversity and structure in a comprehensive sample of New World maize races, the vast majority

TABLE 1. Diversity of the four main clusters of maize races.

Cluster	No. of plants	Gene diversity (SE)	Number of alleles (SE)[a]	Percentage of loci with common allele frequency >0.5[a]	F_{IS}[b]	F_{ST}[c]
Highland Mexican	87	0.814 (0.012) A	14.9 (0.68) A	0.11 A	0.334	0.0291
Tropical lowland	187	0.803 (0.014) A	14.4 (0.69) A	0.17 A	0.333	0.0394
Andean	235	0.706 (0.023) B	12.4 (0.81) B	0.38 B	0.294	0.0268
Northern U.S.	35	0.718 (0.015) B	10.6 (0.53) C	0.34 B	0.444	0.0528

Notes: The four clusters were defined using the software Structure. Standard errors were calculated across loci. Values with different letters differ significantly different from each other based upon pairwise Wilcoxon signed-rank tests.

[a]Mean values from 1000 resamples of 35 plants per cluster. Standard errors (calculated across loci) in parentheses.

[b]Calculated for each cluster separately, treating each as a single population. Based upon bootstrapping 10 000 times across loci, all four F_{IS} estimates were significantly different from zero ($P < 0.01$).

[c]Calculated for each cluster separately, after subdivision into the subclusters identified by Structure (see Table S3.1 in Appendix S3 in Supplemental Data with online version of article). Based upon bootstrapping 10 000 times across loci, all four F_{ST} estimates were significantly different from zero ($P < 0.01$).

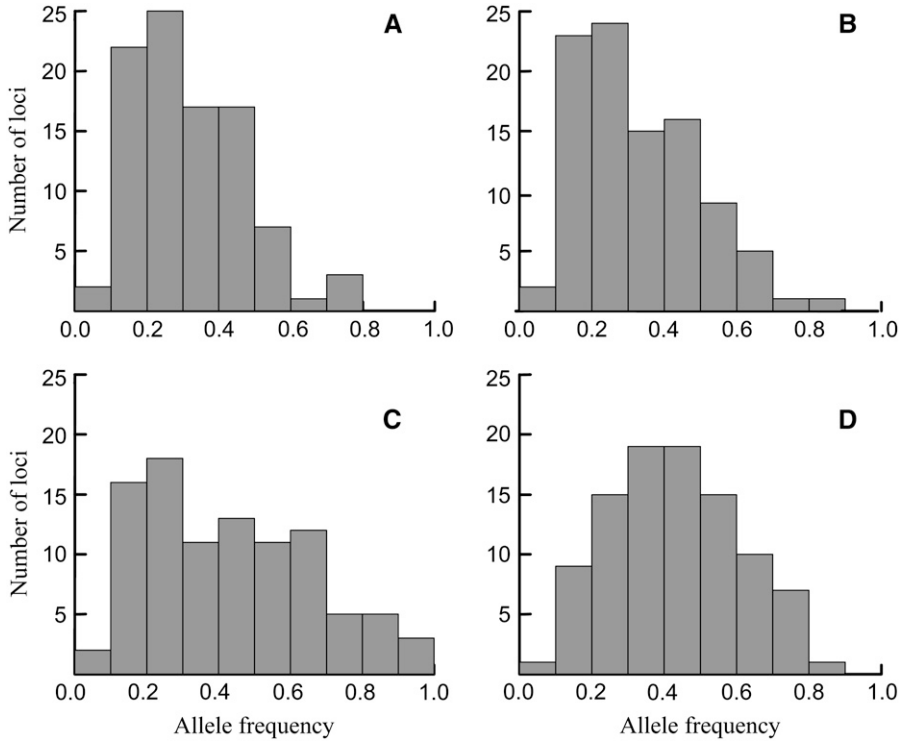

Fig. 5. Histograms of the frequency of the most common allele at 96 microsatellite loci in each of the four main race clusters identified by Structure: (A) highland Mexican, (B) tropical lowland, (C) Andean and (D) northern US.

of which have been in existence since pre-Columbian times. Given that a significant amount of post-Columbian germplasm exchange has undoubtedly occurred (Goodman and Brown, 1988), we used the program Structure in an attempt to recover the sets of contemporary accessions that are most representative of the main clusters of original races. Four main clusters of races were identified, corresponding to highland Mexican, tropical lowland, Andean, and northern US races. However, a large proportion of plants (44%) appeared to have ancestry in more than one of these clusters, which may be a true reflection of admixture between neighboring pre-Columbian clusters or instead be an artifact from the attempt to model discrete populations in the face of continuous gradation of allele frequencies (i.e., isolation by distance, with no obvious boundaries defining adjacent populations or groupings of races) (Pritchard et al., 2000; Rosenberg et al., 2002; Serre and Pääbo, 2004; Barbujani and Belle, 2006). This latter interpretation would seem to be supported by the continuous distribution of maize races and by the clear pattern of isolation by distance observed in this study. However, the same isolation by distance pattern would have been observed if, once the main pre-Columbian race clusters were established, extensive, human-mediated germplasm exchange occurred between them, mainly in the transitional zones between neighboring clusters. Long-distance wind pollination may have also contributed to the blurring of previously established boundaries and thus to the prevalent mixed-ancestry observed here.

TABLE 2. Analyses of molecular variance for maize races of the Americas.

Source of variation	df[a]	Sum of squares	Variance components[b]	Percentage of variation
Among races[c]	309	17733.0	**Va 2.67**	7.6
Among plants[d] within races	654	26697.0	**Vb 8.42**	24.0
Within plants[d]	964	23125.0	**Vc 23.99**	68.4
Total	1927	67555.0	35.1	
Among clusters[e]	3	1862.5	**Va 2.96**	7.5
Among subclusters[f] within clusters	14	1405.5	**Vb 1.27**	3.2
Among plants[g] within subclusters	389	17093.8	**Vc 11.04**	27.7
Within plants[g]	407	9502.5	**Vd 24.55**	61.6%
Total	813	29864.3	39.84	

[a]df, degrees of freedom

[b]Boldfaced values indicate significant differences from zero at $P < 0.0001$.

[c]All 310 races sampled.

[d]All 964 genotyped plants.

[e]Four main clusters identified by Structure analysis.

[f]Eighteen subclusters identified by Structure analysis.

[g]Only plants that were assigned to a subcluster were included (407 plants).

Scenario for maize expansion—The results of our phylogenetic analyses are generally concordant with the scenario of pre-Columbian maize expansion previously proposed by Matsuoka et al. (2002). In this scenario, northwards expansion occurred as follows: maize spread into northern Mexico from its origin of domestication in the Balsas Basin of southwestern Mexico, then into the southwestern US from northern Mexico, and finally into the northern US and Canada from the southwestern US. Our phylogenetic results clearly showed that the northern US races were derived from those of the southwestern US. Our Structure results support this scenario

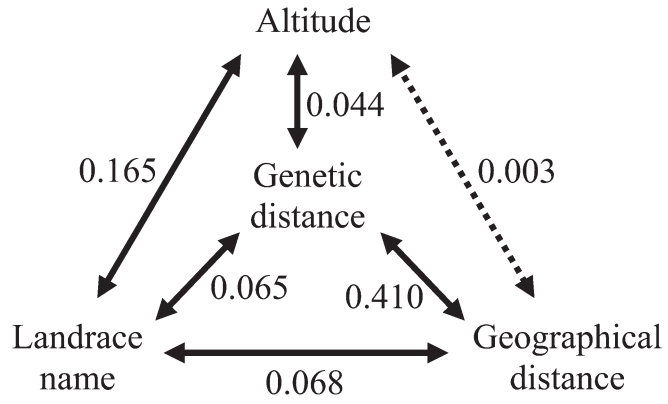

Fig. 6. Mantel test results for correlations between genetic distance, altitude, race, and geographical distance. All correlations were highly significant ($P < 0.001$) except for that between altitude and geographical distance (dashed arrow; $P > 0.05$).

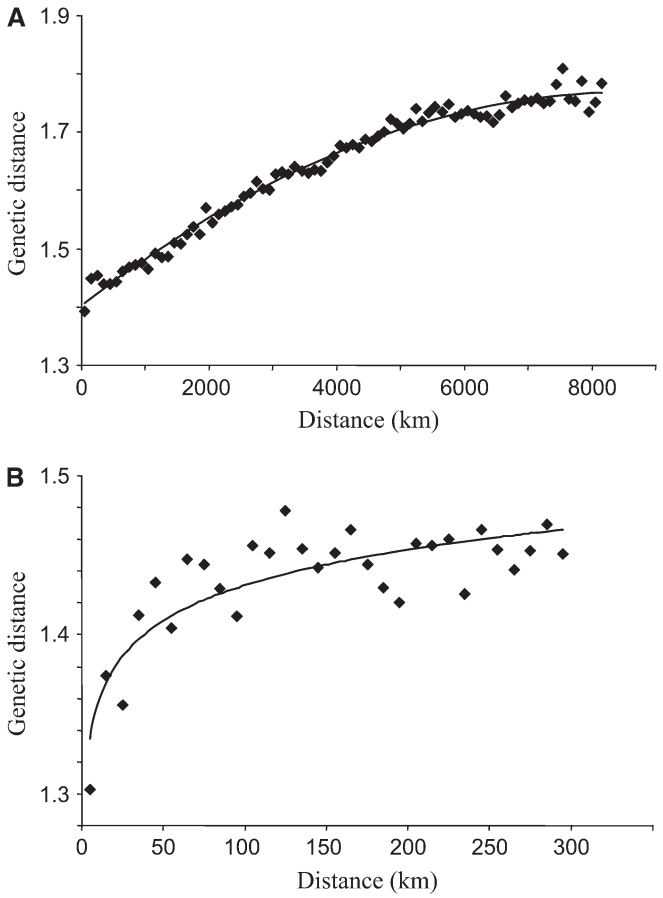

Fig. 7. Relationship between the genetic and geographical distance between plants. For each successive distance class, the average genetic distance (log transform of the proportion of shared alleles) for all pairs of plants falling within that class is plotted. (A) Distance classes of 100 km covering the entire distance spectrum; the polynomial curve that best fits the data ($y = -5 \times 10^{-9}x^2 + 8.5 \times 10^{-5}x + 1.4$; $R^2 = 0.98$) is shown. (B) Finer-scale analysis of plant pairs separated by less than 300 km, with 10 km distance classes; the logarithmic curve that best fits the data ($y = 0.032 \cdot \ln x + 1.28$; $R^2 = 0.71$) is shown.

by indicating that the southwestern US races are intermediate to the highland Mexican and northern US races.

For the pre-Columbian southward expansion of maize, our phylogenetic results support the following scenario. From the highlands of Mexico, maize spread to the lowland tropics of Mexico, then in turn to Guatemala, Colombia, and Venezuela. The Caribbean was then populated from South America via Trinidad and Tobago, and the Andes were populated from Colombia. We uncovered no evidence to suggest that Andean maize descended from the high elevation maize of Guatemala; for example, the multiple accessions sampled of the high elevation Andean race Montaña did not cluster with their suspected Guatemalan counterpart Olotón in the individual-based dendrogram (Fig. 4A), suggesting that adaptation to high elevation had to be acquired de novo in the Andes. The geographically isolated races of central Chile were initially derived from races of the Andes, perhaps in two separate events. Combined interpretation of our Structure and phylogenetic results indicate that lowland middle South America (Bolivia, Argentina, Paraguay, and Uruguay) is the contact zone in which races originating from the Andes have interbred with those from the east coast of South America. This scenario is concordant with that proposed by Freitas et al. (2003), who, based upon the analysis of sequence data from the *alcohol dehydrogenase 2* gene, proposed that maize spread into South America via two routes, a highland route along the Andes and a lowland route along the northeast coast. Middle South America appears to be the meeting ground of these two expansions. Andean origin of some Argentinean races is also supported by the results of Lia et al. (2007), who found that archeological specimens from northwestern Argentina possessed alleles specific to Andean races at three microsatellite loci. Furthermore, Andean genetic signatures in some lowland middle South American races have been observed by McClintock et al. (1981) and by Sánchez G. et al (2007).

To this pre-Columbian historical scenario of maize race colonization, numerous post-Columbian movements must be added. Northern US ancestry is apparent in some Chilean and Argentinean races (such as Dulce Golden Bantam, Dulce Evergreen, Cristalino Norteño and Cateto Sulino Precoce), resulting from the relatively recent introduction to Chile and Argentina of northern US races adapted to high latitudes (Timothy et al., 1961). Partial northern US ancestry in races from Chile was also reported by Dubreuil et al. (2006). In the Structure analysis, races of the southeastern US appeared to be of mixed tropical lowland and northern US ancestry. The most likely scenario is that southeastern US races are a mixture of northern flint germplasm and races imported from the lowland east coast of Mexico and/or from the Caribbean (Doebley et al., 1988; Goodman and Brown, 1988). Furthermore, grouping of most of the southeastern US races sampled herein with those of lowland middle South America in the individual plant phylogenetic analysis most likely resulted from the importation of southeastern US races into Brazil after the Civil War by immigrant farmers from the United States (Brieger et al., 1958; Paterniani and Goodman, 1978; Sánchez G. et al., 2007) and their subsequent introgression into middle South American material. Accessions from the Brazilian and Argentinean races Dente Branco Riograndense, Hickory King and Canario de Ocho group *within* the southeastern US clade in the phylogenetic analysis and thus, of the studied accessions, are the most closely related to the original US imports (see also Sánchez G. et al., 2007). In addition, this genetic analysis confirmed that races with Cubano in their

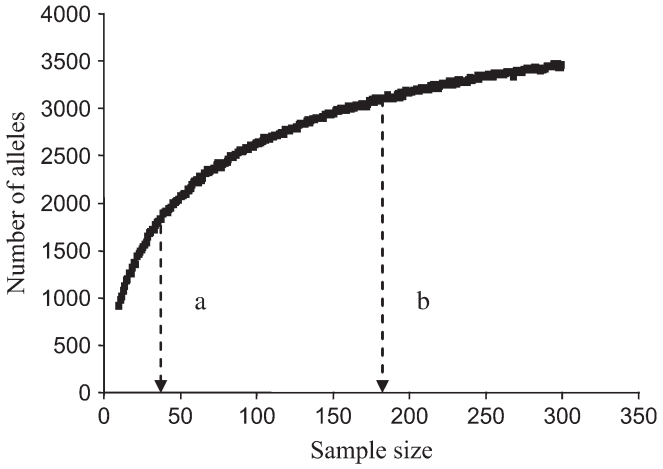

Fig. 8. The maximum number of alleles that can be captured for a given number of sampled accessions, as determined by the line selection algorithm from the computer program PowerMarker (Liu and Muse, 2005). Dotted arrows indicate the minimum number of accessions needed to capture (a) 50% and (b) 80% of the total number of alleles found in this study.

names that are now present in the Andes region were indeed translocated there from Cuba. Finally, the race Alemán from Peru was classified by Structure as belonging to the tropical lowland cluster and was placed by the phylogenetic analysis in the northern South American clade. This race is known to have been brought to Peru from either lowland Mexico or the Caribbean by German settlers (Grobman et al., 1961); though based on only a single Alemán plant, our results support this scenario.

Comparative genetic diversity—Northern US and Andean races contain less genetic diversity than do highland Mexican and tropical lowland races, as measured either by allelic richness, gene diversity, or the average frequency of the most frequent allele at each locus. Similarly, Sánchez G. et al. (2000b) found isozyme allelic diversity to be highest in Mexico and within-race expected heterozygosity to be lowest in northern flint and Andean races. Genetic diversity was likely lost during maize expansion from Mexico into South and North America as a result of serial founder effects. Furthermore, bottlenecks associated with adaptation to new climates and soils may have led to additional, genome-wide reductions of diversity. Maize was first cultivated in the southwestern United States around 1000 BC (Fritz, 1995). Widespread cultivation of maize in the northern United States is thought to have begun much later, in approximately AD 900 (Fritz, 1995). These late introductions may have been partially due to the difficulties in adapting maize first to the dry southwestern environments, then to the short growing seasons and long day lengths of northern latitudes. The northern US cluster was the most divergent because it was highly differentiated from the three other clusters (F_{ST} from 0.116 to 0.175). High divergence of northern flint accessions was previously documented using isozyme data (Doebley et al., 1986). In contrast, genetic diversity levels were not significantly different between the highland Mexican and tropical lowland clusters. However, both genetic diversity and the number of alleles were lower in the tropical lowland cluster vs. the highland Mexican cluster. The observation that genetic diversity is highest in the highlands of Mexico is consistent with the proposed origin of maize in this region (Matsuoka et al., 2002). Interestingly, the expansion of maize into the lowland tropics seems not to have been associated with significant founder effects or selection bottlenecks. It seems either that historical maize introductions into these areas were large enough to mitigate these potential erosive effects on diversity or that subsequent gene flow and new mutations have replenished much of the diversity that was initially lost.

Race name, geographical location, and genetic distance—Two plants sharing the exact same race or cultivar name tend to be more genetically related than two randomly sampled plants, as indicated by the significant Mantel test comparing race name and genetic distance. However, the correlation between race name and genetic distance was weak ($R = 0.058$). This weak correlation was also reflected by the frequent placement of different accessions of the same race into different subclusters by

TABLE 3. Core sets of accessions capturing 50% and 80% of the total allelic diversity found in this study of maize races of the Americas.

Sample size	Number of alleles[a]	Accession numbers
41	1876 (50%)	BOV 755, BOV 968, BOV 968, BOV1071, BOV1132, CAQ 327, CHH 254, CHI 334, CHI 449, CHS 114, DGO 123, ECU 326, ECU 330, ECU 626, GRO 176, GUA 280, GUA 4, GUA 476, JAL 43, MEX 5, MEX 72, MT I, NAY 16, NAY 198, NAY 32, OAX 70, PI213733, PI213801, PI218130, PI218142, PI218148, PI218167, PI311243, PI401754, PIU 2, SON 105, TAM 3, VEN 733, VER 128, YUC 148, ZAC 12
172	3002 (80%)	AGS 7, ANC 181, ANC 184, ARG III, ARG VI, ARQ 22, AYA 22, BOV 396, BOV 438, BOV 712, BOV 743, BOV 755, BOV 992, BOV1000, BOV1132, BOY 453, CAQ 327, CHH 131, CHH 180, CHH 218, CHH 254, CHH 256, CHI 302, CHI 361, CHI 362, CHI 446, CHO 311, CHS 114, CHS 31, CHS 38, CHS 52, CHS 521, CHS 53, CHS 684, CHS 687, CHS 695, COA 36, COL. COMP. M.D., CUN 480, ECU 321, ECU 326, ECU 330, ECU 413, ECU 422, ECU 427, ECU 531, ECU 573, ECU 595, ECU 604, ECU 630, ECU 701, ECU 860, ECU 891, ECU 894, ECU 923, ECU 929, ECU 942, GRO 176, GRO 383, GTO 191, GTO 69, GTO 88, GUA 101, GUA 111, GUA 134, GUA 159, GUA 161, GUA 27, GUA 280, GUA 3, GUA 369, GUA 373, GUA 391, GUA 410, GUA 456, GUA 522, GUA 597, GUA 863, JAL 100, JAL 102, JAL 141, JAL 142, JAL 146, JAL 154, JAL 43, JAL 54, JAL 71, JAL 753, JAL 78, LBQ 17, LBQ 5, LIB 16, LIB 24, LIM 34, LIM 50, LOR 9, MAG 408, MAG 443, MEX 108, MEX 48, MEX 5, MIC 214, MOR 17, MOR 99, MT I, NAR 315, NAR 521, NAY 16, NAY 198, NAY 203, NAY 32, NAY 41, NAY 46, OAX 177, OAX 298, OAX 51, OAX 565, OAX 570, OAX 70, PAG I, PI213731, PI213733, PI213741, PI213774, PI213801, PI213807, PI214279, PI217410, PI217411, PI217480, PI218130, PI218131, PI218133, PI218136, PI218142, PI218148, PI218151, PI218167, PI218187, PI317679, PI401757, PI483087, PIU 115, PIU 92, PUE 109, PUE 27, PUE 552, PUE 591, PUN 6, QOO 39, QRO 2, SIN 61, SON 117, SP XI, TLA 251, URG VI, VEN 409, VEN 442, VEN 445, VEN 481, VEN 529, VEN 604, VEN 736, VEN 760, VEN 843, VEN 874, VER 311, YUC 148, YUC 7, ZAC 12, ZAC 210, ZAC 4

[a]Percentage of total sample of alleles captured in parentheses.

Structure. These results indicate that, although race names are somewhat informative regarding shared ancestry, only a small amount of information is carried in a race name. When maize races were initially classified (Anderson and Cutler, 1942; Wellhausen et al., 1952, 1957; Hatheway, 1957; Roberts et al., 1957; Brieger et al., 1958; Ramírez et al., 1960; Brown, 1960; Grobman et al., 1961; Timothy et al., 1961, 1963; Grant et al., 1963), morphological and ecological similarities were the main considerations. The weak correlation between race name and genetic distance may thus be a reflection of a low correlation between genetic distance at microsatellites and morphological distance. In addition, errors or variance in the attribution of race names (i.e., slightly different names for the same race or the same name for different races) may have also weakened the correlation. In the future, joint analysis of existing morphological, isozyme, and microsatellite data—and possibly new single nucleotide polymorphism and sequence data—may potentially lead to a better classification of maize into more clearly defined races and may shed additional light on its evolutionary and dispersal history (or confirm the results reported here).

In contrast to race name and genetic distance, a much higher correlation was detected between geographical and genetic distance, indicating that there is a strong geographical component to the organization of genetic diversity at the continental scale. This strong geographical component is illustrated in our ancestry maps for the four different clusters (Fig. 3), the high correlation between genetic distance and geographical distance (Fig. 6) and the clear pattern of increasing genetic distance as a function of geographical distance (Fig. 7). This clear geographical pattern undoubtedly resulted from the progressive diffusion of maize culture from Mexico to South and North America after maize domestication. At the broad scale of hundreds to thousands of kilometers (Fig. 7A), genetic distance increases gradually, but steadily. In contrast, at a finer scale of less than 300 km (Fig. 7B), genetic distance between individuals increases dramatically between zero and 50 km, and then only gradually beyond that. Hence, 50 km seems to be, on average, the approximate scale at which local seed exchange occurs among maize farmers, which—together with gene flow via pollen—prevents further differentiation. However, the scale of local germplasm exchange likely varies a great deal among regions, depending on the prevalence of local geographical and cultural barriers. Seed exchange must be far more restricted in the Andes region, with its wildly fluctuating topography, than in the tropical lowland plains. Nevertheless, very low differentiation among race populations at microsatellite loci has been observed at a similar scale in a traditional agricultural system in Oaxaca, Mexico (Pressoir and Berthaud, 2004a). Interestingly, the low microsatellite marker differentiation observed in the Pressoir and Berthaud (2004a) study was associated with strong genetic differentiation at morphological and quantitative traits at the same spatial scale, based upon samples from the same populations (Pressoir and Berthaud, 2004b). These results suggest that farmers play a critical role, via artificial selection for ear characteristics, in maintaining morphological differences between races at the local scale, in the face of extensive gene flow via seed exchange among farmers and neighboring villages. Similar observations of the important role that traditional Mexican farmers play in maintaining the desirable characteristics of cultivated varieties of squash in the face of extensive gene flow were made by Montes-Hernández et al. (2005). These findings may explain the low correlation between genetic distance at microsatellites and race name observed here.

Analysis of molecular variance showed that there is low differentiation between races or between clusters (from 7 to 8%). The vast majority of the genetic diversity lies within races (i.e., among and within accessions of a race). However, the inbreeding coefficient (F_{IS}) is high inside each cluster suggesting finer population structure—i.e., a Walhund effect—within each cluster. Given the large geographical areas occupied by each of the four race clusters, such within-cluster Walhund effects are not surprising: random mating could not be expected to occur at such scales. Moreover, the manner in which seed collections are regenerated, which decreases polymorphism over time, may also have contributed to the high F_{IS} observed; on the other hand, many of the accessions (e.g., those sourced from the National Research Council, or 177 of the 945 accessions), were regenerated only once. In addition, significant heterozygote deficiencies have been observed at the field scale in a traditional farming system and have been attributed to flowering time heterogeneity among plants within a field, leading to assortative mating (Pressoir and Berthaud, 2004a).

Definition of core set sample—We observed a large number of alleles in our full sample of 964 plants, with an average of 39 alleles per locus. We have defined two core set samples of accessions representing the minimum number of accessions needed to capture 50% and 80% of the total number of alleles present in our full sample. Fifty percent of the alleles would be captured by a small sample of only 41 accessions (note, however, that if one wished to capture 50% of the alleles present in the complete gene bank of maize races, and not just in our sample of 964 plants, a much larger sample than 41 accessions would be needed). The tropical lowland and highland Mexican clusters contributed the most to this core set of 41 accessions capturing 50% of our sampled alleles. This is not surprising since these two clusters were the most genetically diverse. Based upon allelic diversity (Table 1), the four clusters, highland Mexican, tropical lowland, Andean and northern US, would be expected to make up 28%, 28%, 24%, and 20% of the accessions in the core sets, respectively; however the actual contributions of the four clusters to the core set of 41 plants were 34%, 32%, 18% and 16%, respectively. Hence, it is clear that the tropical lowland and highland Mexican clusters are overrepresented. This suggests that individual plants that contain combinations of multiple rare alleles across loci are more easily found in more diverse populations. The extent to which maximal capture of rare microsatellite allele variants will correlate with maximal capture of alleles of potential agronomic importance is an open question. Nonetheless, it is clear from this analysis that maize germplasm from highland Mexico and the tropical lowlands contains the most genetic diversity. Liu et al. (2003) have shown that maize inbred line genetic diversity mainly originated from tropical lowland, northern flint and southern dent germplasm. Hence, the high diversity present in races from the highlands of Mexico is underrepresented in maize inbred lines (Liu et al., 2003). Moreover, there is as much diversity in the Andean cluster as in the northern United States. From these results, it is clear that maize diversity is under-utilized in elite inbred lines. Hence, there is vast potential for maize race germplasm to act as a source of novel alleles for the future improvement of elite maize, provided that modern breeding techniques can be used to overcome the susceptibility of highland germplasm to the heat and and fungal pathogens present in lowland environments (Goodman, 2004).

LITERATURE CITED

ANDERSON, E., AND H. C. CUTLER. 1942. Races of *Zea mays*. I. Their recognition and classification. *Annals of the Missouri Botanical Garden* 29: 69–88.

BARBUJANI, G., AND E. M. S. BELLE. 2006. Genomic boundaries between human populations. *Human Heredity* 61: 15–21.

BRETTING, P. K., M. GOODMAN, AND C. W. STUBER. 1990. Isozymatic variation in Guatemalan races of maize. *American Journal of Botany* 77: 211–225.

BRIEGER, F. G., T. A. GURGEL, E. PATERNIANI, A. BLUMENSCHEIN, AND M. R. ALLEONI. 1958. Races of maize in Brazil and other eastern South American countries. National Academy of Sciences, National Research Council, Washington, D.C., USA.

BROWN, W. L. 1960. Races of maize in the West Indies. National Academy of Sciences, National Research Council, Washington, D.C., USA.

BROWN, W. L., AND E. ANDERSON. 1948. The southern dent corns. *Annals of the Missouri Botanical Garden* 35: 255–267.

BUCKLER, E. S., T. PHELPS-DURR, C. KEITH BUCKLER, R. K. DAWE, J. DOEBLEY, AND T. HOLTSFORD. 1999. Meiotic drive of chromosomal knobs reshaped the maize genome. *Genetics* 153: 415–426.

CAMUS-KULANDAIVELU, L., J. B. VEYRIERAS, D. MADUR, V. COMBES, M. FOURMANN, S. BARRAUD, P. DUBREUIL, B. GOUESNARD, D. MANICACCI, AND A. CHARCOSSET. 2006. Maize adaptation to temperate climate: Relationship between population structure and polymorphism in the *Dwarf8* gene. *Genetics* 172: 2449–2463.

CHEVENET, F., C. BRUN, A. L. BAÑULS, B. JACQ, AND R. CHRISTEN. 2006. TreeDyn: Towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics* 7: 439.

DOEBLEY, J., M. GOODMAN, AND C. W. STUBER. 1985. Isoenzyme variation in the races of maize from Mexico. *American Journal of Botany* 72: 629–639.

DOEBLEY, J. F., M. GOODMAN, AND C. W. STUBER. 1986. Exceptional divergence of northern flint corn. *American Journal of Botany* 73: 64–69.

DOEBLEY, J. F., J. D. WENDEL, J. S. C. SMITH, C. W. STUBER, AND M. M. GOODMAN. 1988. The origin of cornbelt maize: The isozyme evidence. *Economic Botany* 42: 120–131.

DUBREUIL, P., M. WARBURTON, M. CHASTANET, D. HOISINGTON, AND A. CHARCOSSET. 2006. More on the introduction of temperate maize into Europe: Large-scale bulk SSR genotyping and new historical elements. *Maydica* 51: 281–291.

EVANNO, G., S. REGNAUT, AND J. GOUDET. 2005. Detecting the number of clusters of individuals using the software Structure: A simulation study. *Molecular Ecology* 14: 2611–2620.

EXCOFFIER, L. 2007. Analysis of population subdivision. *In* D. Balding, M. Bishop, and C. Cannings [eds.] Handbook of statistical genetics, 3rd ed., 980–1020. Wiley, New York, New York, USA.

EXCOFFIER, L., G. LAVAL, AND S. SCHNEIDER. 2005. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* 1: 47–50.

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. 2005. PHYLIP (Phylogeny inference package), version 3.6. Computer program distributed by the author, Department of Genome Sciences, University of Washington, Seattle, Washington, USA.

FREITAS, F. O., G. BENDEL, R. G. ALLABY, AND T. A. BROWN. 2003. DNA from primitive maize races and archaeological remains: Implications for the domestication of maize and its expansion into South America. *Journal of Archaeological Science* 30: 901–908.

FRITZ, G. J. 1995. New dates and data on early agriculture: The legacy of complex hunter-gatherers. *Annals of the Missouri Botanical Garden* 82: 3–15.

FUKUNAGA, K., J. HILL, Y. VIGOUROUX, Y. MATSUOKA, J. SÁNCHEZ G., K. LIU, E.S. BUCKLER, AND J. DOEBLEY. 2005. Genetic diversity and population structure of teosinte. *Genetics* 169: 2241–2254.

GOODMAN, M. M. 2004. Developing temperate inbreds using tropical maize germplasm: Rationale, results, conclusions. *Maydica* 49: 209–219.

GOODMAN, M. M., AND W. L. BROWN. 1988. Races of corn. *In* G. F. Sprague and J. W. Dudley [eds.], Corn and corn improvement, 3rd ed., 33–79. American Society of Agronomy, Madison, Wisconsin, USA.

GRANT, U. J., W. H. HATHEWAY, AND D. H. TIMOTHY. C. CASSALETT D., AND L. M. ROBERTS. 1963. Races of maize in Venezuela. National Academy of Sciences, National Research Council, Washington, D.C., USA.

GROBMAN, A., W. SALHUANA, AND R. SEVILLA. 1961. Races of maize in Peru. National Academy of Sciences, National Research Council, Washington, D.C., USA.

HATHEWAY, W. H. 1957. Races of maize in Cuba. National Academy of Sciences, National Research Council, Washington, D.C., USA.

JAENICKE-DESPRÉS, V., E. S. BUCKLER, B. D. SMITH, M. T. P. GILBERT, A. COOPER, J. DOEBLEY, AND S. PÄÄBO. 2003. Early allelic selection in maize as revealed by ancient DNA. *Science* 302: 1206–1208.

LIA, V. V., V. A. CONFALONIERI, N. RATTO, J. A. C. HERNÁNDEZ, A. M. M. ALZOGARAY, L. POGGIO, AND T. A. BROWN. 2007. Microsatellite typing of ancient maize: Insights into the history of agriculture in southern South America. *Proceedings of the Royal Society, B, Biological Sciences* 274: 545–554.

LIU, J., M. GOODMAN, S. MUSE, AND J. S. SMITH, E. BUCKLER, AND J. DOEBLEY. 2003. Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. *Genetics* 165: 2117–2128.

LIU, J., AND S. V. MUSE. 2005. PowerMarker: An integrated analysis environment for genetic marker analysis. *Bioinformatics (Oxford, England)* 21: 2128–2129.

MCCLINTOCK, B., T.A. KATO Y., AND A. BLUMENSCHEIN. 1981. Chromosome constitution of races of maize. Colegio de postgraduados, Chapingo, Mexico.

MATSUOKA, Y., Y. VIGOUROUX, M. GOODMAN., J. SÁNCHEZ G., E. BUCKLER, AND J. DOEBLEY. 2002. A single domestication for maize shown by multilocus microsatellite genotyping. *Proceedings of the National Academy of Sciences, USA* 99: 6080–6084.

MONTES-HERNÁNDEZ, S., L. C. MERRICK, AND L. E. EGUIARTE. 2005. Maintenance of squash (*Cucurbita* spp.) landrace diversity by farmers' activities in Mexico. *Genetic Resources and Crop Evolution* 52: 697–707.

NEI, M. 1972. Genetic distance between populations. *American Naturalist* 106: 283–291.

PATERNIANI, E., AND M. M. GOODMAN. 1978. Races of maize in Brazil and adjacent areas. CIMMYT, Mexico City, Mexico.

PETIT, R. J., A. EL MOUSADIK, AND O. PONS. 1998. Identifying populations for conservation on the basis of genetic markers. *Conservation Biology* 12: 844–855.

PRITCHARD, J. K., M. STEPHENS, AND P. DONNELLY. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959.

PRESSOIR, G., AND J. BERTHAUD. 2004a. Patterns of population structure in maize races from the central valleys of Oaxaca in Mexico. *Heredity* 92: 88–94.

PRESSOIR, G., AND J. BERTHAUD. 2004b. Population structure and strong divergent selection shape phenotypic diversification in maize races. *Heredity* 92: 95–101.

RAMÍREZ, E. R., D. H. TIMOTHY, E. DÍAZ B., AND U. J. GRANT. 1960. Races of maize in Bolivia. Publication no. 747, National Academy of Sciences, National Research Council, Washington, D.C., USA.

REIF, J., M. L. WARBURTON, X. C. XIA, D. A. HOISINGTON, J. CROSSA, S. TABA, J. MUMINOVIC, M. BOHN, M. FRISCH, AND A. E. MELCHINGER. 2006. Grouping of accessions of Mexican races of maize revisited with SSR markers. *Theoretical and Applied Genetics* 113: 177–185.

ROBERTS, L. M.,, U. J. GRANT. R. RAMÍREZ E., W. H. HATHEWAY, AND D. L. SMITH. 1957. Races of maize in Colombia. National Academy of Sciences, National Research Council, Washington, D.C., USA.

ROSENBERG, N. A., J. K. PRITCHARD, J. L. WEBER, H. M. CANN, K. K. KIDD, L. A. ZHIVOTOVSKY, AND M. W. FELDMAN. 2002. Genetic structure of human populations. *Science* 298: 2381–2385.

SÁNCHEZ G., J. J., M. M. GOODMAN, R. M. BIRD, AND C. W. STUBER. 2006. Isozyme and morphological variation in maize of five Andean countries. *Maydica* 51: 25–42.

SÁNCHEZ G., J. J., M. GOODMAN, AND C. W. STUBER. 2000a. Isozymatic and morphological diversity in the races of maize of Mexico. *Economic Botany* 54: 43–59.

SÁNCHEZ G., J. J., M. GOODMAN, AND C. W. STUBER. 2007. Racial diversity of maize in Brazil and adjacent areas. *Maydica* 52: 13–30.

SÁNCHEZ G., J. J., C. W. STUBER, AND M. GOODMAN. 2000b. Isozymatic diversity in the races of maize of the Americas. *Maydica* 45: 185–203.

SERRE, D., AND S. PÄÄBO. 2004. Evidence for gradients of human genetic diversity within and among continents. *Genome Research* 14: 1679–1685.

SMOUSE, P. E., J. C. LONG, AND R. R. SOKAL. 1986. Multiple regression and correlation extensions of the mantel test of matrix correspondence. *Systematic Zoology* 35: 627–632.

TANKSLEY, S. D., AND S. R. MCCOUCH. 1997. Seed banks and molecular maps: Unlocking genetic potential from the wild. *Science* 277: 1063–1066.

TIMOTHY, D. H., W. H. HATHEWAY, U. J. GRANT, M. TORREGROZA C., D. SARRIA V., AND D. VARELA A. 1963. Races of maize in Ecuador. National Academy of Sciences, National Research Council, Washington, D.C., USA.

TIMOTHY, D. H., B. PEÑA V., AND E. R. RAMÍREZ. 1961. Races of maize in Chile. National Academy of Sciences, National Research Council, Washington, D.C., USA.

WELLHAUSEN, E. J., A. FUENTES O., AND A. HERNÁNDEZ-CORZO. 1957. Races of maize in Central America. National Academy of Sciences, National Research Council, Washington, D.C., USA.

WELLHAUSEN, E. J., L. M. ROBERTS, AND E. HERNÁNDEZ X. 1952. Races of maize in Mexico: Their origin, characteristic and distribution. Bussey Institution of Harvard University, Cambridge, Massachusetts, USA.