DOI: 10.1534/genetics.104.032086

An Analysis of Genetic Diversity Across the Maize Genome Using Microsatellites

Yves Vigouroux,*[,1,2] Sharon Mitchell,[†,1] Yoshihiro Matsuoka,[‡,1] Martha Hamblin,[†] Stephen Kresovich,[†] J. Stephen C. Smith,[§] Jennifer Jaqueth,[§] Oscar S. Smith[§] and John Doebley*[,3]

**Department of Genetics, University of Wisconsin, Madison, Wisconsin 53706, [‡]Fukui Prefectural University, Matsuoka-Cho, Yoshida-gun, Fukui, 910-1195, Japan, [§]Crop Genetics Research and Development, Pioneer Hi-Bred International, Johnson, Iowa 50131 and [†]Institute of Genomic Diversity, Cornell University, Ithaca, New York 14853*

Manuscript received June 5, 2004
Accepted for publication November 30, 2004

ABSTRACT

How domestication bottlenecks and artificial selection shaped the amount and distribution of genetic variation in the genomes of modern crops is poorly understood. We analyzed diversity at 462 simple sequence repeats (SSRs) or microsatellites spread throughout the maize genome and compared the diversity observed at these SSRs in maize to that observed in its wild progenitor, teosinte. The results reveal a modest genome-wide deficit of diversity in maize relative to teosinte. The relative deficit of diversity is less for SSRs with dinucleotide repeat motifs than for SSRs with repeat motifs of more than two nucleotides, suggesting that the former with their higher mutation rate have partially recovered from the domestication bottleneck. We analyzed the relationship between SSR diversity and proximity to QTL for domestication traits and observed no relationship between these factors. However, we did observe a weak, although significant, spatial correlation for diversity statistics among SSRs within 2 cM of one another, suggesting that SSR diversity is weakly patterned across the genome. Twenty-four of 462 SSRs (5%) show some evidence of positive selection in maize under multiple tests. Overall, the pattern of genetic diversity at maize SSRs can be explained largely by a bottleneck effect with a smaller effect from selection.

BETWEEN 5000 and 10,000 years ago, humans domesticated virtually all major crop species used by modern agricultural societies (Smith 2001). This feat was accomplished through artificial selection for traits that improved agronomic qualities. As a result of this process, favorable alleles at loci controlling agronomic traits were brought to fixation in the population during the domestication period. After the initial domestication, the continued practice of selective breeding allowed additional favorable alleles to sweep through the crop species, while diversifying selection in response to the different environments encountered during the geographic expansion of the crop caused regional fixation of distinct favorable alleles. As a consequence of this complex history of selection, only a limited portion of the population contributed to each subsequent generation. Some anticipated consequences are a genome-wide loss of diversity at unselected genes because of the genetic bottleneck effect, a severe reduction in diversity at genes under directional selection during domestication, and artificially high diversity at genes under diversifying selection.

These two processes—selection targeted on agronomic genes and drift due to the domestication bottleneck affecting the entire genome—are the principal factors that influence the amount and distribution of genetic variation in crop genomes as compared to their wild progenitors. Studies on isozymes and gene sequences revealed a general reduction of genetic variation in crops as a result of the domestication bottleneck (Doebley *et al.* 1984; Eyre-Walker *et al.* 1998; Hilton and Gaut 1998); however, these exploratory studies involved relatively few loci and thus the generality of their results needs confirmation. Our knowledge of the impact of selection on diversity in crops is more restricted since very few agronomic genes have been identified and characterized for their level of genetic diversity (Wang *et al.* 1999; Whitt *et al.* 2002; Tenaillon *et al.* 2004). Thus, our present picture of how drift and selection have sculpted the diversity landscape of crop genomes is fragmentary.

To begin to better define genetic diversity in the maize (*Zea mays* ssp. *mays*) genome and to identify the forces that have shaped it, we have constructed a diversity map of the maize genome using microsatellites or simple sequence repeats (SSRs). We scanned the genomes of maize and its close wild relatives, annual teosinte (*Z. mays*

[1]These authors contributed equally to this work.

[2]*Present address:* Diversité et Génomes des Plantes Cultivées, UMR141, Institut de Recherche Pour le Développement, Montpellier, 34394, France.

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

ssp. *huehuetenangensis*, ssp. *mexicana*, and ssp. *parviglumis*), using 462 SSRs. The phylogenetic relationships of these taxa are well known (Doebley 1990; Buckler and Holtsford 1996) and ssp. *parviglumis* has been shown to be the progenitor of maize (Wang *et al.* 1999; Matsuoka *et al.* 2002b). Because of this well-characterized phylogeny, maize and annual teosinte provide a good model for the analysis of the genetic consequences of domestication. The goals of this study are (1) to provide a general picture of genetic diversity for SSRs in maize and teosinte, (2) to determine if there is heterogeneity in diversity among genomic regions, (3) to measure the relative impact of selection *vs.* drift on the observed pattern of diversity, and (4) to assess the degree to which mutation has allowed SSRs to recover diversity lost from the effects of domestication.

MATERIALS AND METHODS

Plant materials: We sampled individual maize plants from a set of 45 landraces covering the entire pre-Columbian range of maize. We also sampled 45 annual teosinte plants representing three wild taxa: *Z. mays* ssp. *huehuetenangensis* (1 plant), ssp. *mexicana* (23 plants) and ssp. *parviglumis* (21 plants). Passport data for the plants are available at www.genetics.org/supplemental (Table S1).

SSRs: We used 462 SSRs, representing a variety of repeat types from dinucleotide to hexanucleotide motifs, distributed throughout the genome. These SSRs were divided in two groups, dinucleotide and "other" repeat SSRs, because the mutation rate for dinucleotide SSRs is higher than that for other SSR types (Vigouroux *et al.* 2002a). Detailed information on the SSRs used in this study including their genetic map position is available at www.genetics.org/supplemental (Table S2). The source of SSRs, whether from expressed sequence tags, known genes, or SSR-enriched genomic libraries, is available at www.maizgdb.org (see also Sharopova *et al.* 2002). SSR genotyping was done on automated sequencers at Cornell University (Ithaca, NY), Pioneer Hi-Bred International (Johnston, IA), and Celera AgGen (Davis, CA), following procedures that have been published elsewhere (Matsuoka *et al.* 2002b).

Statistics: Gene diversity or heterozygosity (H), the number of alleles (N), and F_{st} between maize and teosinte were calculated using the software program Fstat (Goudet 2001). The significance of F_{st} was assessed by 10,000 resamplings of the genotypic data. To measure the relative deficit of gene diversity (GD) in maize *vs.* teosinte, we have defined a parameter $\Delta GD = 1 - (H_M/H_T)$, where H_M and H_T are genetic diversity in maize and teosinte, respectively. If H_M is higher than H_T, then we calculated this parameter as $\Delta GD = (H_T/H_M) - 1$. The relative deficit of the number of alleles is $\Delta allele = 1 - (N_M/N_T)$, where N_M and N_T are the number of alleles in maize and teosinte, respectively. If N_M is higher than N_T, then we calculated this parameter as $\Delta allele = (N_T/N_M) - 1$. These statistics vary between -1 and 1, positive when diversity is higher in teosinte and negative otherwise. The Wilcoxon signed-rank test (W), Kruskal-Wallis test (KW), and Mann-Whitney test (MW) were performed using SYSTAT (SPSS, Chicago).

QTL effects: Prior work has identified a large number of quantitative trait loci (QTL) that differentiate maize and teosinte and can be considered to represent domestication QTL (Doebley and Stec 1993). Positive selection on these QTL during domestication is predicted to cause a reduction in diversity in the surrounding region of the genome. The severity of this reduction at an SSR will be a function of genetic distance (r measured in centimorgans) from the QTL and of the strength of selection (s). The latter is unknown but it is reasonable to consider the effect of the QTL as proportional to s; *i.e.*, QTL of large effect were under stronger selection than those of modest effect. We used the proportion of the variance (V) explained by the individual QTL in the QTL mapping populations as a measure of QTL effect. Thus, for each position (SSR) along a chromosome, we calculated the overall QTL effect (QE) as the sum of V's for the n individual QTL as a function (f) of their distance in centimorgans (r) from the position in question:

$$QE = \sum_{i=1}^{n} V_i \times f(r_i).$$

The relationship between s, r, and diversity statistics (Δallele, ΔGD, or F_{st}) is complicated and there is no known function to describe it. Therefore, we took an *ad hoc* approach. Two different functions (f) were investigated: a linear monotonic decrease $f(r) = 50 - r$ and an exponential decrease $f(r) = e^{-\lambda r}$. For the latter, we used two different values (1 and 5) for λ. The QTL effect is almost zero for $\lambda = 1$ after 10 cM and for $\lambda = 5$ after 2 cM. For each particular function, if $r > 50$ cM, the QTL effect was considered to be zero. Spearman correlation coefficients between QE and diversity statistics (Δallele, ΔGD, or F_{st}) over all SSRs were calculated. Only SSRs placed on the IBM v3 map were tested (www.maizgdb.org).

Spatial analysis: To investigate spatial correlation for the diversity statistics, we calculated the semivariance of F_{st}, Δallele, and ΔGD (Armstrong 1998). The semivariance is one-half the variance of the differences in the value of a statistic between all pairs of points separated by a given distance. Pairs of points close together will show a lower semivariance if they are correlated. The underlying assumption is that the difference between diversity at any two points is a function of the distance between the points. The semivariance (γ) was calculated using the formula

$$\gamma(h) = \frac{1}{2N(h)} \sum_{|x_i - x_j| < h} [Z(x_i) - Z(x_j)]^2,$$

where x_i and x_j are the chromosomal map positions of two SSRs, $Z(x_i)$ and $Z(x_j)$ are the values of their diversity statistics, and $N(h)$ is the number of pairs of SSRs separated by a distance h or less (Armstrong 1998). Three different values of h were investigated: 1, 2, and 5 cM.

Because spatial statistics are based on measures of differences between pairs of SSRs, an unusually small or large value at a given locus may strongly influence the overall results. Hawkins (1980) provides a statistical test to detect outliers by comparing each value $z(x)$ at a location x to neighboring (closest) values on the same chromosome. Let n be the number of neighboring values excluding $z(x)$ and let $z(\bar{x})$ be their arithmetic mean and s the standard deviation of the n values; then

$$\sqrt{\frac{n}{n+1}} \frac{z(x) - z(\bar{x})}{s}$$

follows a t-distribution with $n - 1$ d.f. There is no objective criterion for the sample size n, so we chose the five points that were the closest to the location x. Outliers were excluded at the 95% significance level.

To test if a particular value of the semivariance is significantly different from a random effect, we used permutation tests in which the diversity statistics for the SSRs were random-

ized with respect to chromosomal position. One thousand permuted data sets were generated and the probability of finding a value higher than the observed value for a distance class was then calculated using the distribution of the permuted data.

Test of selection: The Ewens-Watterson test of neutrality enables one to detect deviations from a neutral-equilibrium model as either a deficit or an excess of genetic diversity relative to the number of alleles at a locus (Ewens 1972; Watterson 1978). This test was performed using the program Arlequin (Schneider *et al.* 2000). The probability that an SSR fits the neutral expectation under this test was assessed using both the homozygosity test (P_H) and Slatkin's (1994, 1996) exact test (P_E).

The degree of differentiation between populations at a locus as measured by F_{st} can be used to assess whether SSRs show more differentiation than expected under a purely neutral (drift) model (Bowcock *et al.* 1991; Beaumont and Nichols 1996). We tested whether F_{st} between maize and teosinte at SSRs is greater than expected by the domestication bottleneck effect (drift) alone. To do this, F_{st} was conditioned on the total number of alleles in maize and teosinte to control more effectively for the variable mutation rate among maize SSRs (Vigouroux *et al.* 2002a). Three different mutation models were investigated (see below). We set the 95% confidence limits for this one-tailed test using coalescence simulations that incorporate genetic drift due to the domestication bottleneck (see below). We refer to this as the F_{st} test.

Both selection and drift during domestication are expected to reduce gene diversity in maize relative to teosinte. To ask whether SSRs have less variation in maize relative to teosinte than that expected from drift alone, we compared gene diversity in maize *vs.* teosinte for our observed data with the 95% confidence limits for these parameters established by simulations as a two-tailed test (see below). We refer to this as the GD test.

Simulations: The F_{st} and GD tests ask whether divergence between maize and teosinte or gene diversity in maize relative to teosinte deviates from a neutral model that incorporates the domestication bottleneck. To establish 95% confidence limits for these tests, we performed coalescence simulations (Hudson 1990; see also Vigouroux *et al.* 2002b). The model for the simulations involves a crop (maize) that split at some time in the past from its progenitor (teosinte). The maize population undergoes a "bottleneck" during the domestication period and then expands rapidly to a large size while the progenitor population remains at equilibrium from the time of divergence until the present (Eyre-Walker *et al.* 1998; Hilton and Gaut 1998). A sample size equivalent to our experimental samples of maize and ssp. *parviglumis* was used. Separate topologies for maize and ssp. *parviglumis* were simulated first and then the coalescence times for each node in these topologies were added. The bottleneck in the maize topology was taken into account by rescaling the coalescent times during the bottleneck by the ratio of the effective population size of maize during the bottleneck (N_b) divided by the size after expansion (N_m). The nodes of these two topologies at the time of the split between maize and ssp. *parviglumis* were then treated as a new sample for another simulation to create a single topology combining maize and teosinte.

After a genealogy was simulated, the mutation events were superimposed on it using: (1) the infinite allele model (IAM), under which each mutation creates a new allele (Kimura and Crow 1964); (2) the strict stepwise model (SMM), under which each mutation alters the existing allele by a change of one repeat (Ohta and Kimura 1973); or (3) the generalized stepwise model (GSM), under which the probability of mutation is modeled by a symmetric geometric distribution with a parameter p such that the probability of a mutation of size k during one generation is

$$p(1 - p)^{|k|-1}\mu/2$$

for $k \neq 0$ and a mutation rate of μ (Pritchard *et al.* 1999). For the simulations, the parameter p was estimated to be 0.652 using a value for the variance of the mutation size (σ_m^2) of 3.2 determined from a mutation-accumulation study for maize SSRs (Vigouroux *et al.* 2002a; see also Pritchard *et al.* 1999).

For the simulations, we must estimate the time of divergence of maize and its progenitor, the effective population size of the wild progenitor, the effective population size of maize during the bottleneck and after its expansion, the duration of the bottleneck, and the mutation rate for SSRs. The time of divergence was set at 7500 years (Iltis 1983). The ssp. *parviglumis* effective size was fixed to 40,000 (Vigouroux *et al.* 2002a). The duration of the bottleneck and the effective sizes of maize during and after the bottleneck are unknown, but these parameters are not independent from each other. For estimating the relationship between these parameters, we developed a mathematical model for maize domestication using the GSM (see appendix). Fixing the effective population size of the expanded population of maize to 1 million, we simulated bottlenecks of lengths 100, 200, 500, 1000, and 2500 years and determined their corresponding effective population sizes to be 107, 220, 553, 1117, and 2875. We have used these values for the simulation.

The mutation rate for maize SSRs is variable among loci and the mutation rate for any individual SSR is unknown (Vigouroux *et al.* 2002a). Therefore, we have chosen for each simulation a value of this parameter by the following approach. First, a value for gene diversity (or number of alleles) was picked at random from between 0 and 1 (or between 1 and 51 for number of alleles). Second, the mutation rate that gives this gene diversity (or number of alleles) at equilibrium in ssp. *parviglumis* was calculated and used for simulations. Third, we constrained the mutation rate to be $>5 \times 10^{-7}$ in accordance with empirical data (Vigouroux *et al.* 2002a).

F_{st} (as described in Weir 1996, pp. 181–182), gene diversity, and the total number of alleles for both maize and ssp. *parviglumis* were calculated from the results of 500,000 simulations for each mutation model. This information was then used to estimate the median values and the 95% confidence intervals. As gene diversity is a continuous variable, the expected value of the parameter was calculated using a sliding window of ± 0.0125. To analyze how well the simulated results fit our actual data, we took two approaches. First, we constructed decile curves with the simulated data and calculated the number of actual SSRs lying between two decile curves for the F_{st} by the number of alleles' distribution (Bowcock *et al.* 1991). If the model fits the data perfectly, the number of SSRs lying between two deciles curves should be one-tenth of the total number of SSRs studied. Second, we calculated the mean F_{st} on the basis of the simulation results for a given number of alleles. Then, we used these mean values to calculate an overall expected mean F_{st} for a set of SSRs with the same numbers of alleles as observed in the actual data. We then compared this mean F_{st} for the simulated data with that for the actual data. The same two procedures were used to compare the fit between the actual and simulated data for gene diversity except that the mean expected gene diversity in maize was conditioned on observed gene diversity in ssp. *parviglumis*.

RESULTS

Diversity: Maize possesses less variation at SSRs than does teosinte, whether measured as the number of alleles or as gene diversity (Table 1). Over all SSRs, the average number of alleles is significantly lower in maize

TABLE 1

Diversity and relative diversity loss between teosinte and maize landraces

	N	Teosinte	Maize landraces	Relative loss of diversity
No. of alleles				Δallele
Dinucleotide	163	18.1 (±0.59)	14.8 (±0.61)	0.19 (±0.017)
Other repeats	292	8.1 (±0.24)	5.5 (±0.18)	0.28 (±0.015)
Overall[a]	462	11.8 (±0.35)	9.0 (±0.33)	0.24 (±0.012)
Genetic diversity				ΔGD
Dinucleotide	163	0.87 (±0.008)	0.79 (±0.012)	0.09 (±0.011)
Other repeats	292	0.66 (±0.012)	0.55 (±0.012)	0.15 (±0.017)
Overall[a]	462	0.74 (±0.009)	0.64 (±0.010)	0.12 (±0.012)

The average amounts of allele and gene diversity in teosinte and maize are reported for dinucleotide repeats, for the other repeats, and all the SSRs together with the standard errors in parentheses. The relative losses of diversity in the number of alleles (Δallele) and in gene diversity (ΔGD) are also calculated (see text for details).

[a] Overall includes dinucleotide repeats, other repeats, and seven SSRs with unknown repeat core sequences.

landraces (9.0) than in teosinte (11.8; W test, $P < 0.001$). The relative deficit in allele number or Δallele is 0.24, meaning that maize has 24% fewer alleles than teosinte. Gene diversity is also significantly lower in maize (0.64) as compared to teosinte (0.74; W test, $P < 0.001$) with a ΔGD of 0.12 or a 12% deficit in maize relative to teosinte. The deficit in the number of alleles (24%) is significantly greater than the deficit in gene diversity (12%; W test, $P < 0.001$).

Our prior work on mutation rates for maize SSRs indicated that SSRs with dinucleotide repeat motifs have a much higher mutation rate than SSRs with trinucleotide or larger motifs (here called "other repeat SSRs"; Vigouroux *et al.* 2002a). This difference in mutation rates is reflected in the diversity statistics (Table 1). Dinucleotide SSRs have more alleles than other repeat SSRs both in maize (MW test, $P < 0.001$) and in teosinte (MW test, $P < 0.001$). They also have a higher gene diversity in both maize (MW test, $P < 0.001$) and teosinte (MW test, $P < 0.001$). Therefore, in addition to analyses using all the markers, we performed separate analyses for dinucleotide and other repeat SSRs.

For both dinucleotide and other repeat SSRs, the average number of alleles is higher in teosinte than in maize (W test, $P < 0.001$ and $P < 0.001$, respectively); however, the relative deficit in the number of alleles (Δallele) is greater for other repeat SSRs than for dinucleotide SSRs (MW, $P < 0.001$) (Table 1). Maize shows a relative deficit of 28% for the number of alleles at other repeat SSRs, but a deficit of only 19% for dinucleotide SSRs. Gene diversity exhibits the same trends with a higher diversity in teosinte than in maize for both dinucleotide (W test, $P < 0.001$) and other repeat SSRs (W test, $P < 0.001$), but with ΔGD being greater for other repeat than for dinucleotide SSRs (MW test, $P < 0.001$).

Differentiation: F_{st} between maize and teosinte is low with an average value of 0.071 ± 0.004. Overall, the differentiation between maize and teosinte is highly significant ($P \ll 0.001$). Out of the 462 SSRs, 368 exhibit an F_{st} that is significantly >0 at a noncorrected *P*-value of 0.05. Mean F_{st} is higher (MW, $P < 0.001$) for other repeat SSRs (0.087 ± 0.005) as compared to dinucleotide SSRs (0.044 ± 0.004). There is no difference between dinucleotide and other repeat SSRs in the proportion showing a significant F_{st} (*G*-test = 0.63, $P = 0.43$). F_{is} is 0.38 ± 0.010 for maize and 0.43 ± 0.009 for teosinte. F_{is} is similar for dinucleotide and other repeat SSRs for both maize (MW, $P = 0.62$) and teosinte (MW, $P = 0.13$).

Organization of diversity: *Variability of diversity among chromosomes:* The QTL for plant and inflorescence architecture that differentiate maize and teosinte are mostly found on chromosomes 1–5 (Figure 1; Doebley and Stec 1993). Therefore, if selection on these QTL during domestication caused a severe loss of diversity, one might expect some chromosomal effect on diversity. When all the SSRs are considered, we found no chromosome effect for the parameters ΔGD (KW, $P = 0.38$) and F_{st} (KW, $P = 0.22$), but a significant effect for Δallele (KW, $P = 0.006$). If we considered dinucleotide SSRs (Δallele, KW, $P = 0.11$; ΔGD, KW, $P = 0.12$; F_{st}, KW, $P = 0.83$) and other repeat SSRs (Δallele, KW, $P = 0.08$; ΔGD, KW, $P = 0.40$; F_{st}, KW, $P = 0.37$) separately, there are no significant associations. However, if we combined the two probabilities for Δallele for dinucleotide and other repeat SSRs using Fisher's method for combining probabilities (Sokal and Rohlf 1995), we observe a significant chromosome effect ($P = 0.049$). This result suggests that the chromosome effect is driven by both kinds of repeats. Chromosome 4 has the highest value for Δallele followed by chromosomes 6, 10, 7, 8, 5, 9, 1, 3, and 2 in descending order.

Correlation between diversity and domestication QTL: We

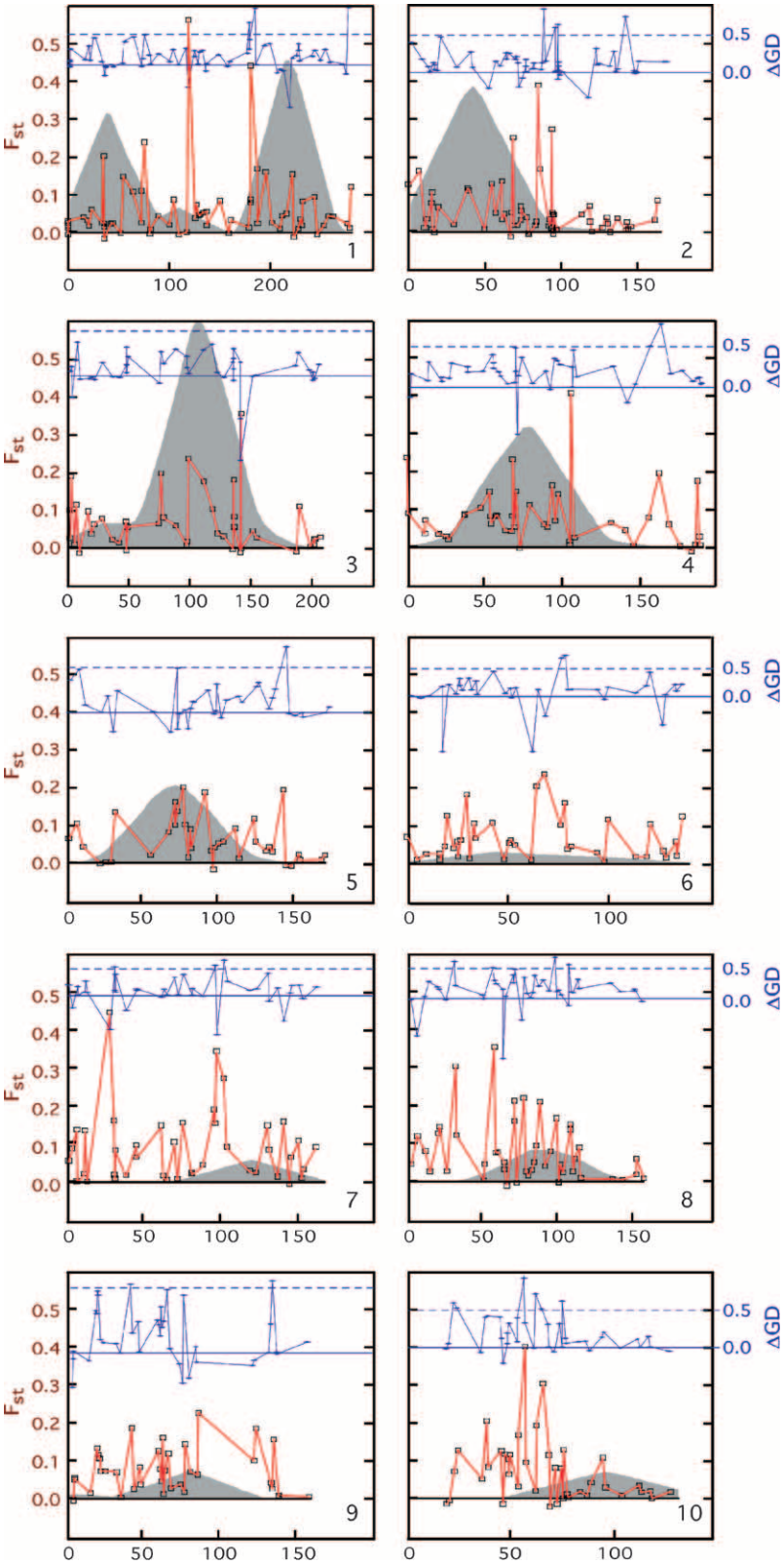

FIGURE 1.—Plot of F_{st} and ΔGD along the genetic map of maize. F_{st} and ΔGD are plotted as a function of the distance in centimorgans along the 10 chromosomes of maize. A representation of domestication QTL effect is shown as a shaded area (see text for details).

can also test if selection on domestication QTL has affected genetic diversity in windows surrounding the individual QTL. If one visually examines the relationship between ΔGD or F_{st} and QTL effect, there is no obvious correlation (Figure 1). At the large-effect QTL region on chromosome 1, neither ΔGD nor F_{st} is particularly large. The same is true for the large-effect QTL regions on chromosomes 2, 3, 4, and 5. Indeed, SSRs with exceptionally large values of F_{st} or ΔGD appear randomly dispersed along the chromosomes.

TABLE 2

Correlation between diversity and the QTL domestication effects

QTL	*N*	ΔAllele	ΔGD	F_{st}
Dinucleotide				
Linear	162	−0.001	0.072	0.194*
Exponential (λ = 1)	162	−0.057	−0.071	0.114
Exponential (λ = 5)	162	−0.063	−0.056	0.146
Other repeats				
Linear	284	−0.063	0.012	0.035
Exponential (λ = 1)	284	−0.032	−0.059	0.048
Exponential (λ = 5)	284	−0.043	−0.103	−0.029
Overall				
Linear	451	−0.042	0.030	0.074
Exponential (λ = 1)	451	−0.066	−0.091	0.038
Exponential (λ = 5)	451	−0.058	−0.097*	0.015

For each individual locus, a QTL effect was calculated, using three different functions of decrease of the effect from the QTL: a linear decrease and two exponential decrease functions (see text for details). *$P < 0.05$.

For a more definitive analysis of the relationship between QTL and SSR diversity, we calculated the correlation between QTL effect and the diversity statistics for the SSRs. If all SSRs are considered together, we observe 1 significant correlation out of 12 between SSR diversity statistics and QTL effect (Table 2). If dinucleotide and other repeat SSRs are analyzed separately, there is also only 1 significant result among 24 tests. We conclude that there is no convincing evidence for a relationship between diversity statistics and QTL effect since single significant tests can readily result by chance alone when doing 24 tests.

Spatial analysis of diversity along the chromosome: In addition to domestication QTL, other spatial factors, such as distance from the centromere, could influence the distribution of diversity. To detect if neighboring SSRs exhibit a similar pattern of diversity, we calculated the semivariance of each of the diversity statistics: Δallele, ΔGD, and F_{st}. If diversity is spatially correlated along the chromosomes, then $\gamma(h)$ for the actual data should be lower than that for a data set obtained by permuting SSRs. Using all SSRs and values of 1, 2, and 5 cM for *h*, we observed significant ($P > 0.95$) values for $\gamma(h)$ for all of the diversity statistics (Table 3). The analysis using only other repeat SSRs gives a similar result. For dinucleotide SSRs, only Δallele and ΔGD show significance, perhaps because of the smaller number of dinucleotide SSRs and corresponding reduced statistical power. Thus, there is evidence that diversity at neighboring SSRs is correlated within recombination distances ranging from 1 to 5 cM. We note that significant spatial correlations are observed only when outlier SSRs were removed from the analysis. Outlier SSRs may result from the variability in mutation rate among SSRs or misplacement of SSRs on the genetic map.

To examine further whether the significant correlations in Table 3 are strictly dependent on the exclusion of outliers, we also calculated $\gamma(h)$ using the *P*-values from the F_{st} test and the GD test for SMM (see below). The use of *P*-values reduces the noise introduced by differences in mutation rates among SSRs. For this analysis, we calculated the odds ratio of the *P*-value as $\ln(p/(1-p))$. Using all the SSRs, we found significant ($P > 0.95$) variogram *P*-values (with outliers, without outliers) from the GD test at 1 cM ($P = 0.981$, $P = 0.969$), 2 cM ($P = 0.997$, $P = 0.995$), but not at 5 cM ($P = 0.93$, $P = 0.804$). For the *P*-values from the F_{st} distribution, we observed significant or near significant associations at 1 cM ($P = 0.944$, $P = 0.904$) and 2 cM ($P = 0.954$, $P = 0.949$), but not at 5 cM ($P = 0.72$, $P = 0.66$). Thus, the exclusion of outliers appears not have biased the observed significant spatial correlation for diversity statistics. Overall, these analyses indicate a significant spatial correlation among SSRs within 2 cM of each other.

Tests of selection: *Ewens-Watterson test:* The Ewens-Watterson test enables one to detect deviations from a neutral-equilibrium model as either a deficit of gene diversity relative to the number of alleles at a locus (below the curve in Figure 2) or an excess of gene diversity (above the curve in Figure 2; Ewens 1972; Watterson 1978). In maize, the number of SSRs showing excess in gene diversity compared to the number of alleles ($P < 0.025$) is 36, and the number showing a deficit in gene diversity ($P > 0.975$) is 12 (supplementary Table S2 at http://www.genetics.org/supplemental/). In teosinte, the number of SSRs showing excess in gene diversity compared to the number of alleles ($P < 0.025$) is 34, and the number showing a deficit in gene diversity ($P > 0.975$) is 5. Maize shows more SSRs with a deficit in gene diversity as expected under selection or a bottleneck.

F_{st} test: The F_{st} test asks if the degree of differentiation at an SSR exceeds neutral expectations. Figure 3 provides a graphical representation of the F_{st} test, showing the medians and upper 95% confidence limits for the SMM, GSM, and IAM established by simulation. The three mutation models give similar results for SSRs with five or fewer alleles; however, for SSRs with more than five alleles, the SMM and GSM have a lower median and 95% confidence limit. To analyze the fit between the simulated model and the observed data set, we calculated the mean of the expected F_{st} for each individual locus given the number of observed alleles. For dinucleotide SSRs, this average is 0.045 (SMM), 0.070 (GSM), and 0.16 (IAM) compared to the observed mean of 0.054. For the other repeats, this average is 0.107 (SMM), 0.138 (GSM), and 0.163 (IAM) compared to the observed mean of 0.097. We also calculated the number of SSRs lying

TABLE 3

Spatial structure of genetic diversity and differentiation

QTL	*N*	Δallele	*P*	*N*	ΔGD	*P*	*N*	F_{st}	*P*
Dinucleotide (cM)									
$h < 1$	27	0.0194	0.98	22	0.0097	0.53	21	0.00103	0.56
$h < 2$	47	0.0354	0.79	39	0.0077	0.81	37	0.00082	0.82
$h < 5$	85	0.0411	0.65	85	0.0068	0.97	79	0.00107	0.71
Other repeats (cM)									
$h < 1$	68	0.0399	0.93	63	0.0349	0.947	58	0.00236	0.99
$h < 2$	118	0.0412	0.98	107	0.0341	0.99	103	0.00325	0.90
$h < 5$	232	0.0435	0.99	228	0.0435	0.94	216	0.00332	0.98
Overall (cM)									
$h < 1$	141	0.0448	0.946	129	0.0279	0.82	125	0.00224	0.96
$h < 2$	252	0.0462	0.98	224	0.0240	0.99	218	0.00243	0.97
$h < 5$	562	0.0509	0.947	509	0.0296	0.86	505	0.00261	0.98

The semivariance of the difference between points separated by a given distance for the three different statistics (Δallele, ΔGD, and F_{st}), the number of pairs of points separated by the given distance (*N*), and the probability that the semivariance is different from a random effect using a resampling procedure (*P*) are given. This analysis was performed with three different distances: 1, 2, and 5 cM.

between consecutive decile curves for each mutation model for both dinucleotide and other repeat SSRs. The IAM does not fit the dinucleotide SSR data because of an excess of SSRs with low F_{st} values ($\chi^2 = 275.4$, $P \ll 0.001$); the GSM and the SMM models are also rejected, but less markedly ($\chi^2 = 29.3$, $P < 0.001$ and $\chi^2 = 21.2$, $P < 0.02$). For the other repeat SSRs, the SMM ($\chi^2 = 7.18$, $P = 0.62$) is not rejected, but the GSM ($\chi^2 = 27.8$, $P < 0.001$) and IAM ($\chi^2 = 63.2$, $P < 0.001$) are rejected. Thus, our actual data best fit the SMM although the fit is not perfect.

With 462 SSRs, the Bonferroni correction threshold would be 0.99989 for the F_{st} test. To test that a locus shows a departure at this *P*-value with good precision

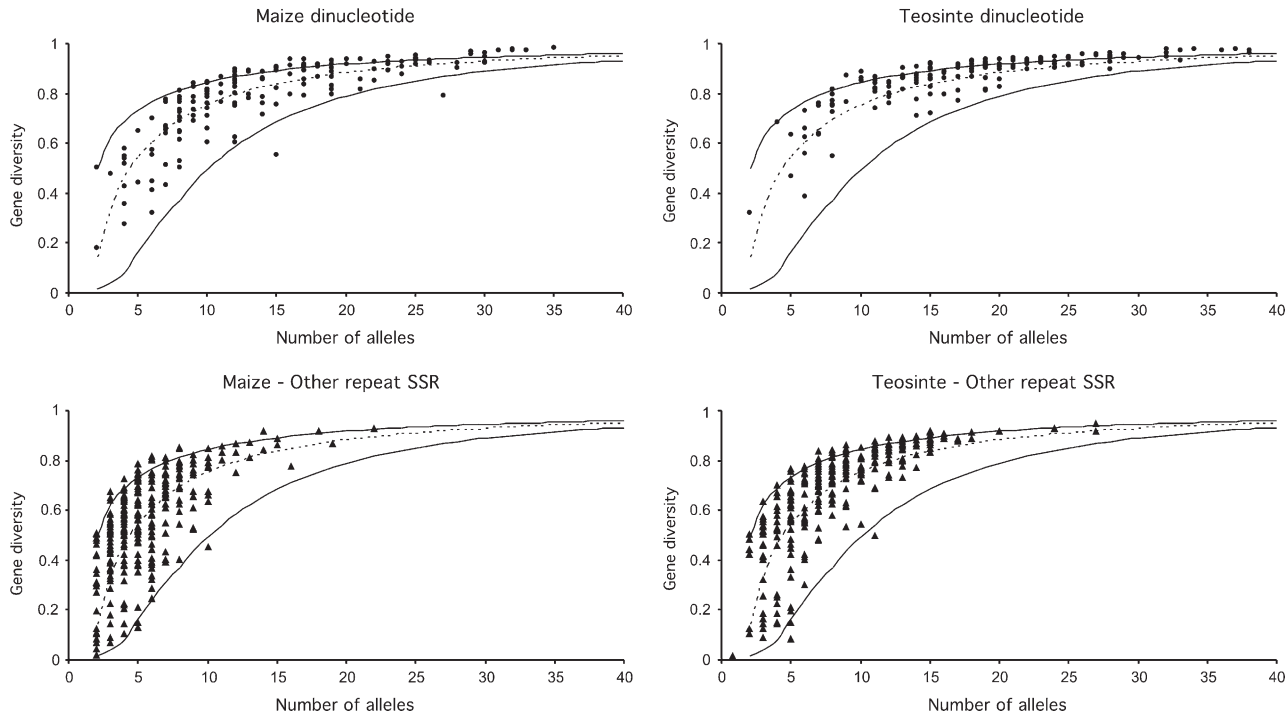

FIGURE 2.—Graphical representation of the Ewens-Watterson test. The 97.5 and 2.5% percentile curves are represented by the solid lines and the expected median value by the dashed line. Circles are dinucleotide and triangles are other repeat SSRs.

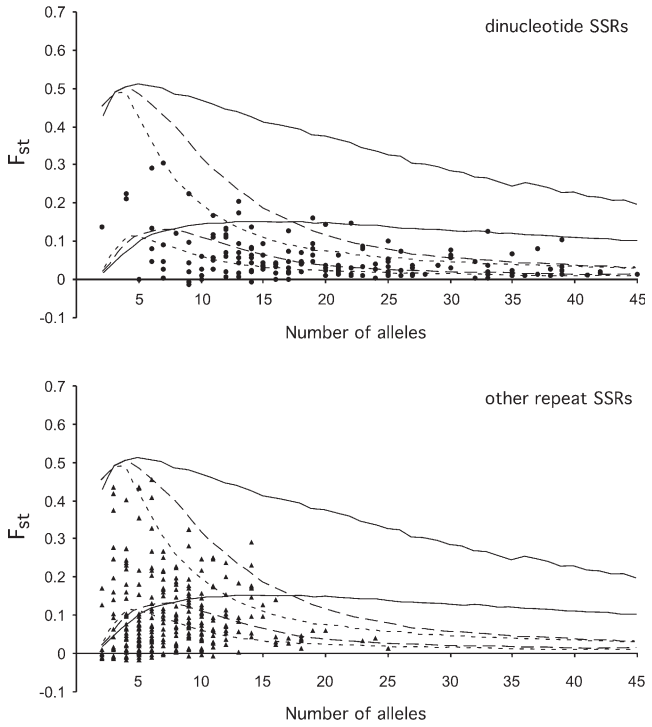

FIGURE 3.—Plot of the F_{st} by the number of alleles. Curves correspond to the 50 and 95% percentiles based on simulated data for three mutation models: an infinite allele model (solid line), the generalized stepwise model (long-dashed line) and a stepwise model (short-dashed line). The plot is presented for dinucleotide SSRs (circles) and other repeat SSRs (triangles).

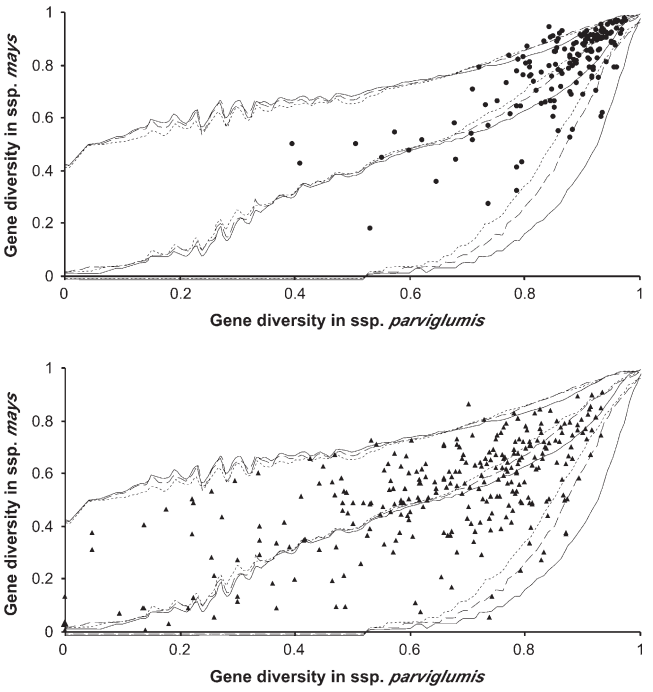

FIGURE 4.—Plot of the gene diversity in maize as compared to the gene diversity in ssp. *parviglumis*. The curves correspond to the 2.5, 50, and 97.5% percentiles based on simulations for three mutation models: an infinite allele model (solid line), the generalized stepwise model (long-dashed line) and a stepwise model (short-dashed line). The plot is presented for dinucleotide SSRs (circles) and other repeat SSRs (triangles).

would require an inordinate number of simulations. So for practical reasons we report here SSRs that exhibit a probability of <0.995 and not the Bonferroni-corrected threshold. Eleven SSRs exhibit higher F_{st} values than expected for the SMM model and zero for both the GSM and IAM at the $P = 0.995$ level. At the $P = 0.95$ level, 46 SSRs are significant for the SMM, 12 for the GSM, and none for the IAM. So with the SMM 10% of the SSRs exhibit a significant value as compared to the 5% expected under a completely neutral distribution.

Gene diversity test: The GD test asks if there has been a greater than expected loss of gene diversity in maize relative to ssp. *parviglumis* given the model for the domestication bottleneck used in the simulations. For all models (IAM, GSM, and SSM), if gene diversity at an SSR in ssp. *parviglumis* is <0.5, then gene diversity in maize can be zero due to loss from the domestication bottleneck alone (Figure 4). To analyze the fit between the simulated model and the observed data, we calculated the mean of the expected gene diversity in maize given the observed gene diversity in teosinte. For dinucleotide SSRs, this average is 0.785 (SMM), 0.768 (GSM), and 0.705 (IAM) compared to the observed 0.787. For the other repeats, this average is 0.541 (SMM), 0.524 (GSM), and 0.495 (IAM) compared to the observed 0.546. We also calculated the number of SSRs lying between consecutive decile curves for each mutation model for both dinucleotide and other repeat SSRs. For dinucleotide SSR data, the IAM ($\chi^2 = 88.7$, $P < 0.001$) is rejected but not the SMM ($\chi^2 = 12.8$, $P = 0.17$) and the GSM ($\chi^2 = 12.7$, $P = 0.18$). For the other repeat SSRs, the SMM ($\chi^2 = 10.1$, $P = 0.35$) and GSM ($\chi^2 = 9.7$, $P = 0.37$) are not rejected, but the IAM is rejected ($\chi^2 = 41.8$, $P < 0.001$). Thus, our data best fit the GSM and SMM, although the fit is not perfect.

For the SMM, 25 SSRs exhibit a significant deficit in diversity in maize relative to teosinte ($P < 0.025$). This represents ~5.4% of the SSRs where only 2.5% (12 SSRs) would be expected by chance. Thus, if the model and parameters used in the simulations are correct, we are likely detecting some SSRs that have reduced diversity because of positive selection during maize domestication or improvement. Fifteen SSRs (3.2%) show a significant excess of diversity in maize ($P > 0.975$) under the SMM where ~12 SSRs would be expected by chance. The expected (12) and observed (15) values are fairly close so there is no compelling evidence for SSRs that are under balancing or diversifying selection in maize.

We summarized the SSRs where two different tests in maize indicate a significant ($P = 0.05$) deviation from neutrality (Table 4). Twenty-nine SSRs in maize show a significant result for multiple tests, 6% of the total number of SSRs. Of these, 24 SSRs or 5% of the 462 show

TABLE 4

Tests of selection

		Sample							
		GD test: maize *vs.* ssp. *parviglumis*			F_{st} test: maize *vs.* ssp. *parviglumis*			Ewens-Watterson: maize	
SSRs	Mutation model:	SMM	GSM	IAM	SMM	GSM	IAM	IAM: p_H	IAM: p_E
bnlg1022		0.0025	0.025		0.995	0.95			
bnlg1046		0.975	0.975	0.975	0.995	0.995			
bnlg1094					0.995	0.95		0.975	
bnlg1182*					0.95			0.9975	0.9975
bnlg1237		0.025			0.95				
bnlg1484		0.0025	0.025		0.995	0.95		0.975	0.975
bnlg1523		0.0025	0.025		0.995	0.995		0.9975	0.9975
bnlg1746		0.9975	0.9975	0.9975				0.025	0.025
bnlg1937		0.0025	0.0025		0.995	0.995			0.975
bnlg426		0.025			0.995	0.95			
mmc0381		0.975						0.025	
nc004		0.025			0.95				
phi050		0.025	0.025	0.025	0.95				
phi96342		0.025			0.995	0.95			
umc1075		0.025			0.95				
umc1078			0.975	0.975					0.025
umc1115		0.025			0.95				
umc1246		0.025			0.95				
umc1299		0.025			0.95				
umc1301		0.025	0.025		0.95				
umc1366		0.975	0.975					0.0025	0.0025
umc1454		0.025	0.025		0.95				
umc1470		0.025			0.95				
umc1675		0.0025	0.025		0.95				
umc1829		0.975	0.975	0.975					0.025
umc1950		0.025			0.95				
umc1970					0.95			0.975	
umc1980		0.0025	0.025		0.95				
umc2031		0.025	0.025		0.995	0.95			

The names of the markers and *P*-values for the selection test are presented for SSRs that have at least two significant probabilities for the tests shown. The F_{st} and the GD tests are based on simulations with the three models of mutation: the infinite allele model (IAM), the generalized stepwise model (GSM), and the stepwise mutation model (SMM). The probability for the Ewens-Watterson test was calculated as the probability obtained by the expected homozygosity (p_H) test and the probability given by an exact test (p_E).

reduced diversity as expected under positive selection. There are similar numbers of dinucleotide and other repeat SSRs with significant tests (Table 4), and these numbers are not significantly different ($G = 3.27$, $P = 0.07$) from a random expectation based on the number in each class of markers in our sample.

DISCUSSION

Genetic diversity and differentiation: Genetic diversity in maize as in other crops has been reduced during domestication as previously shown (Doebley *et al.* 1984; Hilton and Gaut 1998) and further illustrated in this study. For SSRs, maize has 88% of the gene diversity found in teosinte and 76% of the number of alleles. If we divide the SSR data according to the length of the repeat motif, we observe that maize has 91% of gene diversity of teosinte at dinucleotide SSRs and 85% of that at other repeat SSRs. For number of alleles, these values are 81% at dinucleotide SSRs and 72% of that at other repeat SSRs. This deficit of diversity is less than what has been found at the DNA level for *adh1*, 75% (Eyre-Walker *et al.* 1998), or *glb1*, 60% (Hilton and Gaut 1998), as expected since the higher mutation rate for SSRs relative to that for nucleotide substitutions allows SSRs to recover more rapidly from the bottleneck effect (Vigouroux *et al.* 2002a).

We observed a relatively low, although significant, level of differentiation between maize and teosinte ($F_{st} = 0.07$). Since differentiation is driven mostly by drift and both maize and teosinte have large population sizes, the low level of differentiation is not unexpected. Dinu-

cleotide SSRs show a significantly smaller F_{st} value than other repeat SSRs; however, these two types of SSRs exhibit a similar proportion of F_{st} values that are significantly greater than zero. The smaller F_{st} for dinucleotide SSRs occurs because of their higher mutation rate (Vigouroux *et al.* 2002a) and the statistical properties of F_{st}. F_{st} is the function of two probabilities, the probability of identity of two alleles within a population and the probability of identity of two alleles between populations. As the mutation rate increases, the probability of identity within a population decreases and so does the F_{st} value (Weir 1996). This smaller F_{st} value does not mean that the populations are not differentiated, but just illustrates the effect of the mutation rate on F_{st}. The same phenomenon has been observed elsewhere with empirical and simulated data (Balloux *et al.* 2000).

F_{is} is moderately high in both maize (0.38) and teosinte (0.43), but this is likely a function of our sampling strategy. We attempted to maximize the breadth of genetic diversity in our maize and teosinte samples by selecting accessions from maximally divergent geographical locations. This sampling strategy will increase the probability of observing SSRs that have become fixed for alternate alleles in different populations. When multiple plants from single populations are sampled in maize, F_{is} values are much smaller (Labate *et al.* 2003).

Spatial patterning of diversity: A study of the inheritance of domestication traits in maize reported a concentration of QTL on chromosomes 1–5 (Doebley and Stec 1993). This suggests that these chromosomes might have experienced a stronger selective force than chromosomes 6–10 and that there may be heterogeneity among chromosomes in genetic diversity. Nevertheless, no chromosomal effect was detected for either the relative deficit in gene diversity or F_{st}, suggesting a somewhat homogenous genome-wide loss of diversity during domestication (Figure 1, Table 2). The relative deficit of alleles shows some evidence of heterogeneity among chromosomes. Why this effect is observed only for the number of alleles (Δalleles) is unclear. If this effect is due to selection during domestication, it is unlikely that this selection was targeted at the genes (QTL) controlling the differences in plant and inflorescence architecture studied by Doebley and Stec (1993) since the chromosomes that show the most modest losses of alleles (5, 9, 1, 3, and 2) include four of the five chromosomes identified as possessing the largest numbers of QTL.

Diversity and correlation with domestication QTL: We asked whether there is a correlation between the location of domestication QTL and genomic regions of lower genetic diversity as expected if selection during domestication had caused regional losses in diversity. Addressing this question is not straightforward since multiple QTL can be linked in a single region and maize has a complex history. Thus, although the interaction of linkage, selection, and gene diversity has been extensively studied (Maynard Smith and Haigh 1974; Ohta and Kimura 1975; Wiehe and Stephan 1993; Kim and Stephan 2000), no clear models can be applied to maize domestication. For these reasons, we have taken an *ad hoc* approach involving several assumptions: (1) the effect of each domestication QTL on SSR diversity is a decreasing function of the recombination distance to the SSR; (2) the QTL were positively selected; (3) each QTL contributed to the loss of diversity in proportion to the amount of variance it explains (*i.e.*, that selection was stronger for the QTL explaining a higher percentage of the phenotypic variance); and (4) QTL contributed additively to the diversity loss.

Using this approach, we did not observe a significant correlation between QTL effect and loss in the number of alleles (Δallele), gene diversity (ΔGD), or F_{st} (Figure 1). This result can be explained several ways. First, the method we used may not be sensitive enough given the uncertainty of marker positions on the map. Second, we considered here only QTL for morphological traits and not all the potential traits that differentiate teosinte from maize (*e.g.*, seed quality). Third, forces other than directional selection (drift, mutation, diversifying selection) may have created sufficient noise to obscure much of the signal from directional selection. Fourth, none of the SSRs may be sufficiently close to the QTL to have been affected by selection on the QTL. Finally, SSRs used in this study were developed in maize after screening to eliminate invariant SSRs, giving an ascertainment bias since invariant SSRs, which are the most likely candidates for selected SSRs, were excluded from our sample (see Vigouroux *et al.* 2002b).

Diversity correlation between linked SSRs: Selective sweeps or background selection can reduce diversity throughout a chromosomal region (Maynard Smith and Haigh 1974; Charlesworth *et al.* 1993). Therefore, we tested whether linked SSRs are more similar in diversity and we observed multiple significant tests for pairs of SSRs within distances of 2 cM from one another (Table 3).

What mechanisms could produced this correlation? One interpretation is that we are detecting regional variation in the strength of selection during domestication. Where selection was strongest, maize is less diverse (or more differentiated from teosinte) relative to regions that experienced weaker selection. This interpretation, if correct, would appear to contradict prior evidence that the effects of selection on diversity in maize are very narrow (Wang *et al.* 1999) and that that linkage disequilibrium between loci decreases rapidly (Remington *et al.* 2001; Tenaillon *et al.* 2001). Another interpretation may be that there is some bias in the data (or in the parameters) that creates a correlation among neighboring SSRs. For example, if there are regions of high *vs.* low recombination and if recombination is correlated with SSR mutation rate (see Tenaillon *et al.* 2001), then a statistic like F_{st} that is influenced by the mutation rate could show a spatial correlation in the

absence of any effect from selection during domestication.

Tests of neutrality: *Simulating SSR evolution in maize:* To test whether an SSR exhibits a nonneutral pattern of variation, one needs to know the neutral distribution against which the observed data can be compared. To compute such a distribution, we have used coalescent simulations that incorporate the domestication bottleneck. These simulations were performed using three different models for microsatellite evolution: IAM, SMM, and GSM. The simulations are also based on estimates of the current effective population size of maize, the duration of the bottleneck, and the population size of maize during the bottleneck (Eyre-Walker *et al.* 1998; Vigouroux *et al.* 2002a). Error in these estimates could bias the results. Nevertheless, this approach has the advantage of clearly specifying the model used and takes into account some aspects of maize history, although it does not include more complex features like population structure.

We examined the fit between our actual data and the simulated data and found that the mean gene diversity and F_{st} values from the simulated data were closest to the actual data when the simulations were based on the SMM as opposed to the IAM and GSM. Similarly, the distributions of the gene diversity and F_{st} values for our actual data were closest to the simulated distributions when the simulations were based on the SMM. Overall, the SMM fit the actual data in three of the four tests performed. Nevertheless, the fit is not exact and the results of the simulations differ from expectations based on our prior empirical work. Notably, our prior work on SSR mutation rates (Vigouroux *et al.* 2002a) indicates that dinucleotide SSRs should best fit the GSM, while a study of sequence diversity at other repeat SSRs (Matsuoka *et al.* 2002a) suggests that the IAM might provide the best model for this class of SSR. Other factors not incorporated into the simulations such as population structure or directional evolution (Vigouroux *et al.* 2003) could be responsible for the imperfect fit between the actual and simulated data. Therefore, caution is advised in interpreting the simulation results and the tests of neutrality based upon them.

F_{st} and GD tests: We performed two tests of nonneutral evolution for which the expected distribution of the test statistic was determined using coalescent simulations. For the F_{st} test, 46 SSRs or 10% of the 462 SSRs exhibited a higher F_{st} value between maize and teosinte than expected under the SMM at the $P = 0.05$ significance level or twice the expected number (23) under purely neutral evolution (Table S4 at http://www.genetics.org/supplemental/). For the GD test, 25 SSRs or 5.4% of the 462 SSRs exhibit a deficit in diversity relative to teosinte under the SMM at the $P = 0.025$ significance level or twice the expected number (12) under purely neutral evolution (Table S4). This excess of loci with significant F_{st} or ΔGD values suggests that some of these SSRs (or sites closely linked to them) may have been under selection during maize domestication. These loci merit further investigation by DNA sequence analysis to better assess whether they have indeed experienced past selection.

Ewens-Watterson test: We have also investigated the influence of selection on diversity by analyzing individual SSRs for evidence of nonneutral evolution using the Ewens-Watterson test. A large number of SSRs (34 in teosinte and 36 in maize) exhibit excess gene diversity relative to the number of alleles (Figure 2, Table S4). This result may indicate balancing (diversifying) selection or population subdivision (Kreitman 2000). For teosinte, population subdivision is a likely explanation because our sample includes three different clusters, ssp. *parviglumis*, ssp. *mexicana*, and ssp. *huehuetenangensis*, which are highly structured (Matsuoka *et al.* 2002b). Similarly, our maize sample was chosen to maximize the geographic regions represented and does not represent a single Hardy-Weinberg population, an assumption of the Ewens-Watterson test.

In maize 12 SSRs (2.6%) exhibit a deficit in gene diversity relative to the number of alleles as expected under positive selection or a bottleneck (Figure 2, Table S4). This is about the number of significant tests expected by chance alone given the significance threshold of $P = 0.975$ for the two-tailed Ewens-Watterson test. Thus, this test did not enable us to identify any likely targets of selection during maize domestication. In a previous article, we identified 7 of 39 maize SSRs with a deficit in gene diversity relative to the number of alleles using the Ewens-Watterson test (Vigouroux *et al.* 2002b). However, in this prior work, we biased our choice of SSRs to enrich the sample for ones that were likely targets of selection. The failure to identify nonneutral SSRs with the Ewens-Watterson test in the present analysis could also be influenced by ascertainment bias. Since we studied only SSRs that were polymorphic in maize and could thus be placed on the maize genetic map, we systematically excluded low-diversity (invariant) SSRs that are the most likely targets of selection.

Perspective: Our results enable us to make some tentative interpretations concerning the forces that have sculpted SSR diversity across the maize genome. First, we infer that mutation has allowed dinucleotide SSRs with their high mutation rates (10^{-3}–10^{-4}) to partially recover from the loss of diversity during maize domestication. We make this inference since ΔGD for these SSRs is only 9% as compared to 15% for other repeat SSRs, which have a lower mutation rate ($\sim 10^{-5}$) (Vigouroux *et al.* 2002a). Similarly, we infer that other repeat SSRs have also made a partial, although weaker, recovery since ΔGD for these loci is still smaller than the ΔGD of 33% for nucleotide substitutions that have even a lower mutation rate ($\sim 10^{-9}$; White and Doebley 1999). Nevertheless, since SSR gene diversity remains lower in maize than in teosinte at both dinucleotide

and other repeat SSRs, we conclude that new mutation over the ~5000 years since the end of the bottleneck has not produced a complete recovery. Thus, SSR diversity can provide some insights into the relative roles of drift and selection as well.

Given that SSRs show reduced diversity in maize relative to teosinte, we can ask what were the relative roles of drift and selection in producing this reduction. Our data do not allow an unequivocal answer to this question, but they can be used to suggest that drift was the dominant force. First, the results of our coalescent simulations indicate that diversity at the vast majority of SSRs can be explained by a simple model that incorporates the domestication bottleneck (drift), thereby obviating the need to infer selection. Similarly, we observed no correlation between the chromosomal position of domestication QTL and diversity as expected if selection coupled with hitchhiking had caused strong regional reductions in diversity.

Even if drift during the domestication bottleneck is the major factor influencing SSR diversity in maize, we ask whether some SSRs have reduced diversity because of selection during maize domestication. A conservative approach for identifying SSRs that were likely targets of selection is to consider only SSRs that exhibit significant results with two different neutrality tests in maize. Taking this approach we identified 29 of the 462 SSRs that we consider the best candidates for selected SSRs (Table 4). Under a complete independence of the three tests, the probability to observe at least two tests significant at a 0.05 level is 0.00725. So, under neutrality, the expected number of SSRs with two significant tests for 462 SSRs analyzed is 3.3 compared to 29 actually observed. However, the tests are not completely independent from each other, so the number 29 is somewhat of an upper limit of the number of selected SSRs under the neutral models considered.

Of the 29 SSRs with two or three significant tests, 24 show evidence for positive selection during maize domestication as either a higher than expected ΔGD value or a Ewens-Watterson test indicating a deficit of gene diversity as compared to the number of alleles. Thus, an average value of 5% of the total number of SSRs may have experienced positive selection during maize domestication. This value of 5% "selected" loci may underestimate the actual number of genes that have experienced selection for several reasons. For example, we analyzed only loci that were known to be polymorphic in maize and thereby excluded invariant SSRs, the class most likely to have experienced past selection. Similarly, some SSRs that experienced selection during maize domestication some 9000 years ago may have recovered their loss diversity because of the high mutation rate for SSRs and thereby give nonsignificant neutrality test results. Nevertheless, this 5% value provides a rough first estimate of the portion of the maize genome that was under positive selection during maize domestication.

We thank Montgomery Slatkin and Jody Hey for advice on the mathematical and simulation models. We thank Marit Haug for technical assistance and Major Goodman and Jesus Sanchez for help in obtaining seeds. This work is supported by National Science Foundation grant DBI-0096033.

LITERATURE CITED

Armstrong, M., 1998 *Basic Linear Geostatistics*. Springer-Verlag, Berlin/Heidelberg, Germany.

Balloux, F., H. Brunner, N. Lugon-Moulin, J. Hausser and J. Goudet, 2000 Microsatellites can be misleading: an empirical and simulation study. Evolution **54:** 1415–1422.

Beaumont, M. A., and R. A. Nichols, 1996 Evaluating loci for use in the genetic analysis of population structure. Proc. R. Soc. Lond. Ser. B **263:** 1619–1626.

Bowcock, A. M., J. R. Kidd, J. L. Mountain, J. M. Hebert, L. Carotenuto *et al.*, 1991 Drift, admixture, and selection in human evolution: a study with DNA polymorphism. Proc. Natl. Acad. Sci. USA **88:** 839–843.

Buckler, E. S., and T. P. Holtsford, 1996 *Zea* systematics: ribosomal ITS evidence. Mol. Biol. Evol. **13:** 612–622.

Charlesworth, B., M. T. Morgan and D. Charlesworth, 1993 The effects of deleterious mutations on neutral molecular variation. Genetics **134:** 1289–1303.

Doebley, J. F., 1990 Molecular systematics of *Zea* (Gramineae). Maydica **35:** 143–150.

Doebley, J., and A. Stec, 1993 Inheritance of the morphological differences between maize and teosinte: comparison of results for two F2 populations. Genetics **134:** 559–570.

Doebley, J. F., M. M. Goodman and C. W. Stuber, 1984 Isoenzymatic variation in *Zea* (Gramineae). Syst. Bot. **9:** 203–218.

Eyre-Walker, A., R. L. Gaut, H. Hilton, D. L. Feldman and B. Gaut, 1998 Investigation of the bottleneck leading to the domestication of maize. Proc. Natl. Acad. Sci. USA **95:** 4441–4446.

Ewens, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87–112.

Goudet, J., 2001 *FSTAT, a Program to Estimate and Test Gene Diversities and Fixation Indices*, Version 2.9.3 (http://www.unil.ch/izea/softwares/fstat.html).

Hawkins, D. M., 1980 *Identification of Outliers*. Chapman & Hall, London.

Hilton, H., and B. S. Gaut, 1998 Speciation and domestication in maize and its wild relatives: evidence from the *globuline-1* gene. Genetics **150:** 863–872.

Hudson, R. R., 1990 Gene, genealogy and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–44.

Iltis, H. H., 1983 From teosinte to maize: the catastrophic sexual transmutation. Science **222:** 886–894.

Kim, Y., and W. Stephan, 2000 Joint effects of genetic hitchhiking and background selection on neutral variation. Genetics **155:** 1415–1427.

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

Kreitman, M., 2000 Methods to detect selection in populations with applications to the human. Annu. Rev. Genomics Hum. Genet. **1:** 539–559.

Labate, J., K. R. Lamkey, S. E. Mitchell, S. Kresovich, H. Sullivan *et al.*, 2003 Molecular and historical aspects of corn belt dent diversity. Crop Sci. **43:** 80–91.

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

Maynard Smith, J., and J. Haigh, 1974 The hitchhiking effect of a favorable gene. Genet. Res. **23:** 23–35.

Ohta, T., and M. Kimura, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22:** 201–204.

Ohta, T., and M. Kimura, 1975 The effect of selected linked loci on heterozygosity of neutral alleles (the hitch-hiking effect). Genet. Res. **25:** 313–326.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun and M. W. Feldman, 1999 Population growth of human Y chromosomes: a study of Y microsatellites. Mol. Biol. Evol. **16:** 1791–1798.

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.

Schneider, S., D. Roessli and L. Excoffier, 2000 *Arlequin Ver. 2.000, a Software for Population Genetics Data Analysis.* Genetics and Biometry, University of Geneva, Geneva.

Sharopova, N., M. McMullen, L. Schultz, S. Schroeder, H. Sanchez-Villeda *et al.*, 2002 Development and mapping of SSR markers for maize. Plant Mol. Biol. **48:** 463–481.

Slatkin, M., 1994 An exact test for neutrality based on the Ewens sampling distribution. Genet. Res. **64:** 71–74.

Slatkin, M., 1995 Hitchhiking and associative overdominance at a microsatellite locus. Mol. Biol. Evol. **12:** 473–480.

Slatkin, M., 1996 A correction to the exact test based on the Ewens sampling distribution. Genet. Res. **68:** 259–260.

Smith, B. D., 2001 Documenting plant domestication: the consilience of biological and archaeological approaches. Proc. Natl. Acad. Sci. USA **98:** 1324–1326.

Sokal, R. R., and F. J. Rohlf, 1995 *Biometry: The Principles and Practice of Statistics in Biological Research.* W. H. Freeman, New York.

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.

Tenaillon, M. I., J. U'Ren, O. Tenaillon and B. S. Gaut, 2004 Selection versus demography: a multilocus investigation of the domestication process in maize. Mol. Biol. Evol. **21:** 1214–1225.

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

Vigouroux, Y., M. McMullen, C. T. Hittinger, K. Houchins, L. Schulz *et al.*, 2002b Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. Proc. Natl. Acad. Sci. USA **99:** 9650–9655.

Vigouroux, Y., Y. Matsuoka and J. Doebley, 2003 Directional evolution for microsatellite size in maize. Mol. Biol. Evol. **20:** 1480–1483.

Wang, R. L., A. Stec, J. Hey, L. Lukens and J. Doebley, 1999 The limits of selection during maize domestication. Nature **398:** 236–239.

Watterson, G. A., 1978 The homozygosity test of neutrality. Genetics **88:** 405–417.

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

White, S., and J. Doebley, 1999 The molecular evolution of *terminal ear1*, a regulatory gene in the genus Zea. Genetics **153:** 1455–1462.

Whitt, S., L. M. Wilson, M. Tenaillon, B. S. Gaut and E. Buckler, 2002 Genetic diversity and selection in the maize starch pathway. Proc. Natl. Acad. Sci. USA **99:** 12922–12962.

Wiehe, T. H. E., and W. Stephan, 1993 Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster.* Mol. Biol. Evol. **10:** 842–854.

Communicating editor: S. R. McCouch

APPENDIX: MATHEMATICAL MODEL FOR THE MAIZE DOMESTICATION BOTTLENECK

For an SSR that follows a generalized stepwise model, the recursion equation for the variance in allele size as a function of drift and mutation has been derived by Slatkin (1995) as

$$\sigma_a^2(t+1) = \left(1 - \frac{1}{2N}\right)\sigma_a^2(t) + \mu\sigma_m^2, \tag{A1}$$

where N is the effective population size, $\mu\sigma_m^2$ is the effective mutation rate, and $\sigma_a^2(t)$ is the variance in allele size at generation t. Assuming discrete nonoverlapping generations, genetic drift will reduce the variance in allele size by a factor of $(1 - 1/2N)$ and mutation will increase the variance by $\mu\sigma_m^2$. Equation A1 can be rewritten as

$$\sigma_a^2(t+1) - 2N\mu\sigma_m^2 = \left(1 - \frac{1}{2N}\right)(\sigma_a^2(t) - 2N\mu\sigma_m^2). \tag{A2}$$

We can extend Equation A2 to model the loss of variation during the bottleneck period (T_b), if we assume no gene flow between ssp. *mays* and ssp. *parviglumis* and an effective population size during the bottleneck of N_b. If σ_0^2 is the variance in allele size in the ancestral population at the beginning of the bottleneck, then the variance in allele size at the end of the bottleneck is

$$\sigma_b^2 = \left(1 - \frac{1}{2N_b}\right)^{T_b}\sigma_0^2 + 2N_b\mu\sigma_m^2\left(1 - \left(1 - \frac{1}{2N_b}\right)^{T_b}\right). \tag{A3}$$

Similarly, by extension of Equation A2, we find that the variance in allele size for the maize population (σ_{maize}^2) T generations after the bottleneck is

$$\sigma_{maize}^2 = \left(1 - \frac{1}{2N_b}\right)^{T_b}\left(1 - \frac{1}{2N}\right)^{T}\sigma_0^2 + 2N_b\mu\sigma_m^2\left(1 - \left(1 - \frac{1}{2N_b}\right)^{T_b}\right)\left(1 - \frac{1}{2N}\right)^{T} + 2N\mu\sigma_m^2\left(1 - \left(1 - \frac{1}{2N}\right)^{T}\right), \tag{A4}$$

assuming an effective size of N during this period. For large N and N_b, this formula can be approximated by

$$\sigma_{maize}^2 \approx e^{-T_b/2N_b - T/2N}\sigma_0^2 + 2N_b\mu\sigma_m^2(1 - e^{-T_b/2N_b})e^{-T/2N} + 2N\mu\sigma_m^2(1 - e^{-T/2N}). \tag{A5}$$

Assuming that the ancestral population was at equilibrium and had the same effective size as ssp. *parviglumis* today,

we can use σ^2_{parvi} as an estimate for σ^2_0. Knowing σ^2_m, N_{maize}, T, and T_b, we can estimate the effective population size of maize during the bottleneck (N_b).

The effective mutation rate for 33 dinucleotide SSRs in maize ($\mu\sigma^2_m$) was estimated using mutation-accumulation studies to be 8.8×10^{-4} (Vigouroux *et al.* 2002a). The mean variance for ssp. *parviglumis* ($\overline{\sigma^2_{parvi}}$) and maize ($\overline{\sigma^2_{maize}}$) over 33 dinucleotide SSRs was estimated using the data from Matsuoka *et al.* (2002b) as 23.5 and 26.8, respectively. The effective population size of the expanded maize population after the bottleneck in a range from 10^5 to 10^9 has only a small effect on the estimated size during the bottleneck (Eyre-Walker *et al.* 1998 and data not shown), so we have considered only a large effective population size of 1 million for maize after the expansion.

With these values for the parameters, we can estimate N_b for different values of T_b using Equation A5. Archaeological information indicates that the domestication bottleneck was probably within the range of a few hundred to 2000 years (Smith 2001). Therefore, we calculated the effective size for bottlenecks of 100, 200, 500, 1000, and 2500 years in duration and obtained values for N_b of 107, 220, 553, 1117, and 2875, respectively. These values are in good agreement with previous independent estimates using DNA sequence polymorphism (Hilton and Gaut 1998).