DOI: 10.1534/genetics.107.085506

The Role of Regulatory Genes During Maize Domestication: Evidence From Nucleotide Polymorphism and Gene Expression

Qiong Zhao,*[,1,2] Anne-Céline Thuillet,*[,1,3] Nathan K. Uhlmann,[†] Allison Weber,* J. Antoni Rafalski,[†] Stephen M. Allen,[†] Scott Tingey[†] and John Doebley*[,4]

**Laboratory of Genetics, University of Wisconsin, Madison, Wisconsin 53706 and* [†]*DuPont Crop Genetics Research, Dupont Experimental Station, Wilmington, Delaware 19880*

Manuscript received December 7, 2007
Accepted for publication February 5, 2008

ABSTRACT

We investigated DNA sequence variation in 72 candidate genes in maize landraces and the wild ancestor of maize, teosinte. The candidate genes were chosen because they exhibit very low sequence diversity among maize inbreds and have sequence homology to known regulatory genes. We observed signatures of selection in 17 candidate genes, indicating that they were potential targets of artificial selection during domestication. In addition, 21 candidate genes were identified as potential targets of natural selection in teosinte. A comparison of the proportion of selected genes between our regulatory genes and genes unfiltered for their potential function (but also with very low sequence diversity among maize inbreds) provided some weak evidence that regulatory genes are overrepresented among selected genes. We detected no significant association between the positions of genes identified as potential targets of selection during domestication and quantitative trait loci (QTL) responsible for maize domestication traits. However, a subset of these genes, those identified by sequence homology as kinase/phosphatase genes, significantly cluster with the domestication QTL. We also analyzed expression profiles of genes in distinct maize tissues and observed that domestication genes are expressed on average at a significantly higher level than neutral genes in reproductive organs, including kernels.

REGULATORY genes play an important role in development by controlling the expression of downstream structural or regulatory genes. It has been suggested that changes in function or expression of regulatory genes may be associated with the diversification of plant morphology (Doebley and Lukens 1998; Purugganan 1998, 2000). Two regulatory genes controlling differences in plant morphology between maize and teosinte, *teosinte branched1* (*tb1*) and *teosinte glume architecture1* (*tga1*), have been identified through quantitative trait locus (QTL) mapping (Doebley *et al.* 1997; Wang *et al.* 2005). Both of these regulatory genes are responsible for major morphology changes that occurred during the domestication of maize. An increase in the expression of *tb1* led to reduced branching in maize (Doebley *et al.* 1997), whereas a change in the function of the *tga1* protein appears to be responsible for reducing the size of the casing around the kernel in teosinte (Wang *et al.* 2005). In other crops, genes that control domestication traits were also revealed to encode regulatory genes, including the tomato gene *fw2.2* affecting fruit weight (Frary *et al.* 2000), the rice shattering gene (Li *et al.* 2006), and the *Q* gene in wheat (Simons *et al.* 2006). This list suggests that regulatory genes may have been important targets of selection during crop domestication.

To identify other maize genes that were targets of selection during domestication, approaches based on molecular population genetics have been employed (Wright and Gaut 2005). Evidence of selection can be detected either by standard tests of the neutral equilibrium model or by a coalescence-simulation (CS)-based test. The coalescence-simulation-based test assays whether the relative loss of nucleotide diversity in maize as compared to teosinte is too large to be accounted for by a domestication bottleneck alone such that selection can be inferred. The Hudson–Kreitman–Aguadé (HKA) test, a standard neutrality test, assays whether the amount of nucleotide diversity in the gene of interest is significantly lower than the amount of nucleotide diversity in neutral genes in maize. Application of these tests provided evidence that the domestication genes *tb1* and *tga1* were both targets of selection during domestication (Wang *et al.* 1999, 2005; Clark *et al.* 2004).

Recently, large-scale genomic screens using molecular population genetics methods have identified a long list of genes that were possible targets of selection during maize domestication (Vigouroux *et al.* 2002;

[1] These authors contributed equally to this work.

[2] *Present address:* Department of Biostatistics, University of Washington, Seattle, WA 98195.

[3] *Present address:* Centre IRD de Montepellier, 911 av. Agropolis, 34394 Montpellier, Cedex 5, France.

[4] *Corresponding author:* Laboratory of Genetics, University of Wisconsin, 425 Henry Mall, Madison, WI 53706. E-mail: jdoebley@wisc.edu.

Wright *et al.* 2005; Yamasaki *et al.* 2005). An initial study found evidence for selection during domestication at 10 loci by screening simple sequence repeats located in 501 maize genes (Vigouroux *et al.* 2002). A subsequent study using single nucleotide polymorphism (SNP) markers and a sample of 774 genes found 30 putative domestication or crop improvement genes (Wright *et al.* 2005). Eight genes with strong evidence of selection during domestication were identified by a third study, which examined genes with zero diversity in 14 maize inbred lines (Yamasaki *et al.* 2005). In total, these studies have identified 48 loci that may have been targets of selection during maize domestication and subsequent improvement.

In this study, we have taken an approach similar to that used by Vigouroux *et al.* (2002) and Yamasaki *et al.* (2005) to investigate candidate genes for signatures of selection associated with maize domestication. Similar to these studies, our candidate gene pool consists of genes with very low genetic diversity in maize inbred lines. Unlike these previous studies, we filtered our candidate gene sample to include only those with sequence homology to known regulatory genes. These include DNA-binding transcription factors, receptor kinases, regulators of RNA metabolism, and components in signal transduction pathways. Our candidate gene sample consisted of 72 expressed sequence tags (ESTs) identified as putative regulatory genes.

Here we report that 17 of our 72 candidate genes (23.6%) exhibit evidence that they were targets of selection during domestication. An additional 21 of our 72 candidate genes (29.2%) were identified as potential targets of selection in teosinte. By comparing our results with those from another study, we conclude that there is minimal evidence that regulatory genes are overrepresented among genes that show evidence of selection. When the genetic map positions of our candidate domestication genes were tested for association with previously mapped QTL responsible for maize domestication traits, we found no evidence that our candidate domestication genes are clustered near domestication QTL. However, map positions from a subset of the 17 candidate domestication genes, those with sequence homology to known kinases and phosphatases, significantly colocalize with known domestication QTL. Finally, by examining expression profiles of genes in distinct maize plant tissues, we observed that candidate domestication genes tend to be more highly expressed relative to neutral genes in kernel and other reproductive tissues as opposed to vegetative tissues where no significant difference was observed.

MATERIALS AND METHODS

Plant materials and sequence data: Nucleotide diversity statistics for a set of sequence alignments for 6995 ESTs and 600 maize lines were made available by Dupont Crop Genetics. Among the 6995 ESTs, a subset of 390 had low genetic diversity ($0 < H_D < 0.05$, where H_D is the haplotype diversity). To determine whether these 390 ESTs shared homology to known classes of regulatory genes, we queried the Entrez protein database using BLASTX (http://www.ncbi.nlm.nih.gov/BLAST). By this means, we identified 72 putative regulatory genes among the 390 with low diversity. For comparison to these 72 candidate genes, we randomly selected 47 additional ESTs from the total set of 6995 ESTs (excluding candidates) to serve as a control set. This set of controls allows us to ask whether genes that are prescreening for both low diversity in maize inbreds and putative regulatory function are more likely to be targets of selection than random genes.

For diversity analysis, we collected DNA sequences from 16 maize landraces and 16 teosinte (*Zea mays* ssp. *parviglumis*) individuals for each of these 119 ESTs (supplemental Table 1). As previously described (Tenaillon *et al.* 2001), this geographically diverse sample of maize landraces represents the genetic diversity present in the maize population before modern breeding efforts. The 16 different teosintes were chosen on the basis of geographic criteria and cover the entire natural distribution of *Z. mays* ssp. *parviglumis.* Three accessions of *Tripsacum dactyloides*, which belong to the sister genus of Zea, were used as outgroup individuals for some analyses (supplemental Table 1). *T. dactyloides* alleles were successfully isolated for 43 of the 72 candidate genes and 22 of the 47 control genes.

Two pairs of PCR primers, including a nested pair, were designed to amplify each EST. Often the targeted segment of the gene included the 3′-untranslated region and a portion of the open reading frame. To facilitate PCR product sequencing, the internal forward and reverse primers were equipped with T3 (5′-aattaaccctcactaaaggg-3′) and T7 (5′-gtaatacgact cactatagggc-3′) 5′-tails. For each locus, an initial PCR reaction was performed using the outer primer set under the following conditions: 95° for 5 min, followed by 24 cycles of 94° for 20 sec, 55° for 30 sec, 72° for 2 min, and a final step of 72° for 10 min. The reaction products were diluted 10-fold with TE buffer and used for a subsequent round of PCR with the nested primer set under the same PCR conditions as described above. The products from this round of PCR were then used for DNA sequencing in both directions, using T3 and T7 primers and a standard protocol (Applied Biosystems, Foster City, CA) on an ABI 3700 DNA sequencer.

The forward and reverse DNA sequences from each individual were assembled using Sequencher software (Gene Codes). Individual sequences from the maize landraces, teosinte, and outgroup individuals were then manually aligned using BioEdit software (Hall 1999). Since our teosinte individuals are partial inbreds, sites where base calls were ambiguous due to potential heterozygosity were coded as "N." Unique single-base-pair variants (singletons) were double checked by manually inspecting their raw chromatogram peaks.

Tests for neutrality: Molecular population genetics statistics were estimated separately for maize landraces and teosinte individuals using DnaSP (Rozas *et al.* 2003). Nucleotide polymorphism (θ) (Watterson 1975) and nucleotide diversity (π) (Tajima 1983) were calculated on the basis of all sites. Estimates of the population recombination rate ($4Nc$, where N is the effective population size and c is the recombination rate per base pair per generation) (Hudson 1987) were also calculated using all sites. The HKA test (Hudson *et al.* 1987) for neutrality was also performed. For the HKA test, *T. dactyloides* was used as the outgroup. Eleven neutral loci (*adh1*, *an1*, *asg75*, *bz2*, *csu1138*, *csu1171*, *csu381*, *csu1132*, *fus6*, *glb1*, and *umc128*) (Eyre-Walker *et al.* 1998; Hilton and Gaut 1998; Tenaillon *et al.* 2001) were used for HKA tests involving maize landraces, while a smaller set of neutral loci (*adh1*, *glb1*, *bz2*, *csu1132*, and *csu1171*) (Tenaillon *et al.* 2004) were used in HKA tests involving teosinte. For each HKA test, an overall χ^2 value was obtained by summing the individual χ^2 values calculated using each neutral locus. This overall χ^2 value was then used to obtain an overall *P*-value.

Coalescence-simulation analysis for selection: For each candidate and control gene, a coalescence-simulation-based test was used to determine if the gene was a potential target of selection during domestication. We used a modified version of the standard coalescence procedure (Hudson *et al.* 1987) that incorporated the domestication bottleneck as previously described (Eyre-Walker *et al.* 1998). All parameters in the model were assigned to previously established values (Eyre-Walker *et al.* 1998; Tenaillon *et al.* 2004). The severity of the bottleneck (k) was defined as a function of the population size during the bottleneck (N_b) and the duration of the bottleneck (d) such that $k = N_b/d$. Using sequence data from 44 neutral genes, the best multilocus estimate of k was found to be 2.0 using methods previously described (Tenaillon *et al.* 2004). To estimate k, we used the number of segregating sites (S) as the summary statistic and d was set equal to 1000. Finally, k values ranging from 0.5 to 5 (in increments of 0.1) were explored.

We used the coalescence model described above to test for selection in 68 candidate genes. Four of the 72 candidate genes were excluded from analysis because no polymorphism in teosinte was observed in which case the test cannot be performed. For each of the 68 candidate genes, 10,000 simulations were conducted. A gene was considered to be a potential target of selection during domestication if S_{maize} was $<97.5\%$ of the S_{simul} values.

Expression analysis: We obtained the expression pattern of candidate and control genes using information from a Massively Parallel Signature Sequencing (MPSS) database (Brenner *et al.* 2000). This database includes 109 tissue libraries generated using the maize inbred line B73 (supplemental Table 2). The MPSS method utilizes a 17-mer sequence tag that is generated using the 3′-most *Dpn*II site in a given cDNA. The abundance of the sequence tag is then measured and used to infer the relative abundance of the corresponding gene transcript (Brenner *et al.* 2000). The accuracy of the expression profiles obtained using the MPSS method was confirmed by comparing these results with previously reported expression patterns for several genes (supplemental Materials and Methods; supplemental Figure 1; supplemental Figure 2). Transcript abundance was recorded in parts per million. Signals were considered as background noise if lower than an arbitrary cutoff of 5 ppm. Using the methods described above, we were able to obtain expression profiles for a total of 66 genes (27 control and 39 candidate genes).

Several analyses including principal component analysis (PCA), permutation *t*-tests, and analysis of variance (ANOVA) were used to characterize any overall pattern in the expression profiles. Using the statistical computing package R, PCA was conducted with transcript abundance transformed to a logarithmic scale. Permutation *t*-tests were used to identify pairwise differences in expression levels between three tissue types (vegetative, kernel, and nonkernel reproductive tissues; supplemental Table 2) or between gene classes (neutral genes *vs.* genes selected during domestication). The effect of interaction between tissue type and gene class on the abundance of expressed transcripts was assessed using ANOVA. For both permutation *t*-tests and ANOVA, signals were transformed to logarithmic scale as in PCA analysis. For the ANOVA, we fit a linear mixed model using an R module (Bates 2007). In our model, gene class and tissue type were considered fixed effects while individual genes and libraries were considered random effects. The significance of the interaction term between tissue type and gene class was determined by comparing the fit of two models, one with the interaction term (model 1) and another model without the interaction term (model 2).

TABLE 1

Sequence statistics of 47 control genes and 72 candidate genes

	Control genes				Candidate genes			
	N[a]	*L*[b]	*S*[c]	θ[d]	*N*[a]	*L*[b]	*S*[c]	θ[d]
Maize landraces	13.5	524	11.3	0.0070	12.8	540	2.9	0.0018
Teosinte	11	524	14.7	0.0099	12.5	540	8.1	0.0052

[a] Average no. of sequences in the alignment.

[b] Average length of alignments, excluding gaps.

[c] Average no. of segregating sites (SNPs) in the alignments.

[d] Average amount of nucleotide polymorphism (Watterson's estimator of population mutation parameter).

RESULTS

Nucleotide diversity in maize and teosinte: First, we compared sequence diversity between the control and candidate genes. The average sequence lengths and number of individuals were similar for the control and candidate genes (Table 1). We estimated θ (Watterson 1975) for the control genes and observed that maize landraces retained 70.7% ($r = \theta_{\text{maize}}/\theta_{\text{teosinte}}$) of the genetic diversity found in teosinte (Table 1). This value is similar to that found in previous studies that reported that maize retained 57–80% of the genetic diversity found in teosinte (Tenaillon *et al.* 2004; Wright *et al.* 2005). Thus, our set of neutral or control genes is consistent with such genes in prior studies.

Nucleotide diversity was significantly lower in both teosinte and maize when measured in the candidate genes as opposed to the control genes (Table 1; supplemental Table 3). For both maize landraces and teosinte, the average number of segregating sites (S) was significantly lower in candidate genes as evaluated by the Mann–Whitney (MW) test (maize landraces, $P < 0.001$; teosinte, $P < 0.001$). The proportion of nucleotide diversity maintained in the maize landraces as compared to teosinte was also lower, $\sim$35% as measured by θ. This calculation excludes four genes with no polymorphism in teosinte (PZC07071, PZC08281, PZC11007, and PZC15464). The greater loss of diversity in maize as compared to teosinte for candidate genes was statistically significant (MW test: $P < 0.001$). This suggests that, overall, the candidate genes have lost more diversity than a random sample of genes from the maize genome.

Statistical tests for neutrality: The HKA test was conducted on a subset of 43 candidate and 22 control genes for which outgroup sequences were obtained. For each gene, the HKA test was performed separately on maize landraces and teosinte. The HKA test is based on the theoretical prediction that under neutrality the ratio of diversity within a species to divergence between this species and an outgroup should be equivalent for all genes in the genome. If this ratio was significantly smaller for a candidate gene than for the neutral genes, then selection was inferred. We considered a gene as a potential target of selection during domestication if the HKA test was significant in maize landraces but not in teosinte. If the HKA test was significant in teosinte, we considered the gene as a target of selection in teosinte. Using the criteria above, we identified three control genes and 14 candidate genes as putative domestication genes (Tables 2 and 3). We also identified four control genes and 21 candidate genes as putative targets of selection in teosinte (Tables 2 and 3).

In contrast to the HKA test, the CS-based test does not require sequence from an outgroup species and thus the majority of the candidate genes could be tested. The CS test asks if the loss of diversity in maize as compared to teosinte is too great to be accounted for by the domestication bottleneck alone. If so, then selection during domestication is inferred. We were able to test all but four of the candidate genes for selection during domestication with the CS test. The four excluded candidate genes had no polymorphism in teosinte. Ten candidate genes demonstrated a greater deficiency of polymorphism in maize than expected from the domestication bottleneck alone and thus were identified as possible domestication genes (Table 3). Among these 10 putative domestication genes, 7 were also identified as putative domestication genes by the HKA test; the remaining 3 genes were not tested with the HKA test because no outgroup sequence was available (Table 3).

Comparison of candidate gene results with previous studies: To assess whether regulatory genes were a major target of selection, we compared the proportion of putative selected genes in our candidate gene pool with that found in a previous study. Yamasaki *et al.* (2005) examined 35 candidate genes that had no diversity in maize inbred lines using a strategy similar to ours. However, their study did not filter candidate genes in regard to function, while we specifically chose candidate genes with sequence homology to putative regulatory genes. To assess the role of regulatory genes in selection, we compared the results from these two sets of candidate genes.

To determine whether any differences between the two sets of candidates could be attributed to a sampling bias, we compared the sequence statistics of the two groups (Table 4). On average, similar numbers of maize landraces and teosintes were amplified for genes in both studies. Although our regulatory gene alignments are almost twice as long as their unfiltered candidate genes, the number of haplotypes (h) and average nucleotide diversity (π) are similar between the two studies for the maize landraces (MW test, $P = 0.59$ for h and $P = 0.53$ for π) as well as for teosinte (MW test, $P = 0.73$ for h and $P = 0.15$ for π).

TABLE 2

Results of the HKA and CS tests for 47 control genes

	P-values from HKA test			*P*-values from CS test	
Gene	Maize	Teosinte	Selection status	Maize *vs.* teosinte	Selection status
DX414418	0.854	0.947	—	0.5808	—
DX414429	NA	NA	NA	0.2656	—
DX414430	0.999	0.991	—	0.8844	—
DX414436	NA	NA	NA	0.3988	—
DX414440	0.995	0.923	—	0.477	—
DX414414	NA	NA	NA	0.493	—
DX414415	NA	NA	NA	0.0222[a]	—
DX414416	NA	NA	NA	0.9152	—
DX414417	0.667	0.984	—	0.2052	—
DX414419	0.993	0.851	—	0.5938	—
DX414420	NA	NA	NA	0.6868	—
DX414421	NA	NA	NA	0.3494	—
DX414422	<0.001	<0.001	Teosinte	0.5796	—
DX414423	NA	NA	NA	0.4126	—
DX414424	NA	NA	NA	0.2282	—
DX414425	0.069	0.795	—	0.5048	—
DX414426	NA	NA	NA	0.7636	—
DX414427	NA	NA	NA	0.686	—
DX414428	0.007	0.204	Domestication	0.5134	—
DX414431	0.106	0.533	—	0.328	—
DX414432	0.148	0.004	Teosinte	0.2992	—
DX414433	0.985	0.510	—	0.12	—
DX414434	0.994	0.760	—	0.297	—
DX414435	0.990	0.957	—	0.4858	—
DX414437	0.950	0.489	—	0.7664	—
DX414438	<0.001	0.003	Teosinte	0.1178	—
DX414439	NA	NA	NA	0.1874	—
DX414441	NA	NA	NA	0.5098	—
DX414442	0.866	0.146	—	0.2058	—
DX414443	0.927	0.958	—	0.6444	—
DX414444	NA	NA	NA	0.4756	—
DX414445	NA	NA	NA	0.3782	—
DX414446	NA	NA	NA	0.269	—
DX414447	0.019	0.663	Domestication	0.0986	—
DX414448	0.703	0.110	—	0.5714	—
DX414449	0.918	0.965	—	0.7206	—
DX414450	NA	NA	NA	0.1196	—
DX414451	0.002	0.897	Domestication	0.0844	—
DX414452	NA	NA	NA	0.2632	—
DX414453	NA	NA	NA	0.3416	—
DX414454	NA	NA	NA	0.8848	—
DX414455	NA	NA	NA	0.3094	—
DX414456	NA	NA	NA	0.709	—
DX414457	NA	NA	NA	0.9542	—
DX414458	0.818	0.000	Teosinte	0.0014[a]	—
DX414459	NA	NA	NA	0.2504	—
DX414413	NA	NA	NA	0.9282	—

NA indicates that an outgroup was not available or that the test was not applicable due to the lack of segregating sites in teosinte. A dash (—) denotes that there was no evidence of selection in either maize or teosinte. Domestication genes and teosinte-selected genes are designated as domestication and teosinte, respectively.

[a] Indicates that a significant *P*-value was generated due to an excess of polymorphism in maize relative to teosinte.

TABLE 3

Results of the HKA and CS tests for 72 candidate genes

	P-values in HKA test			*P*-values in CS test	
Gene	Maize	Teosinte	Selection status	Maize *vs.* teosinte	Selection status
DX414517	0.001	0.021	Teosinte	0.4528	—
DX414490	<0.001	<0.001	Teosinte	0.4104	—
DX414527	<0.001	0.554	Domestication	0.0274	Domestication
DX414473	NA	NA	NA	0.01	Domestication
DX414474	0.110	0.037	Teosinte	0.4382	—
DX414475	<0.001	0.094	Domestication	0.2662	—
DX414510	<0.001	0.931	Domestication	0.0274	Domestication
DX414511	0.891	0.040	Teosinte	0.1306	—
DX414512	<0.001	<0.001	Teosinte	0.983	—
DX414478	<0.001	0.002	Teosinte	0.972	—
DX414479	NA	NA	NA	0.9552	—
DX414542	<0.001	0.001	Teosinte	0.1362	—
DX414480	NA	NA	NA	0.4184	—
DX414513	NA	NA	NA	0.5828	—
DX414481	<0.001	0.006	Teosinte	0.3838	—
DX414482	NA	NA	NA	0.4502	—
DX414484	NA	NA	NA	0.5374	—
DX414515	0.661	0.161	—	0.486	—
DX414516	<0.001	0.823	Domestication	0.2468	—
DX414486	<0.001	0.141	Domestication	0.6552	—
DX414518	0.556	0.510	—	0.6838	—
DX414519	NA	NA	NA	NA	NA
DX414520	NA	NA	NA	0.0116	Domestication
DX414487	NA	NA	NA	0.3448	—
DX414488	<0.001	<0.001	Teosinte	0.8944	—
DX414521	<0.001	<0.001	Teosinte	0.9888	—
DX414522	<0.001	0.185	Domestication	0.3056	—
DX414523	0.086	0.195	—	0.9012	—
DX414489	NA	NA	NA	0.8992	—
DX414524	0.652	0.035	Teosinte	0.3998	—
DX414525	0.204	0.736	—	0.6658	—
DX414544	NA	NA	NA	NA	NA
DX414526	0.938	0.445	—	0.4818	—
DX414492	NA	NA	NA	0.3908	—
DX414528	0.127	0.039	Teosinte	0.684	—
DX414530	NA	NA	NA	0.2924	—
DX414531	<0.001	0.953	Domestication	0.0038	Domestication
DX414532	<0.001	0.662	Domestication	0.0058	Domestication
DX414533	NA	NA	NA	0.0756	—
DX414493	0.018	0.024	Teosinte	0.5532	—
DX414494	<0.001	0.149	Domestication	0.1004	—
DX414534	NA	NA	NA	0.8372	—
DX414495	NA	NA	NA	0.6226	—
DX414460	NA	NA	NA	0.1496	—
DX414496	NA	NA	NA	0.8966	—
DX414497	NA	NA	NA	0.5572	—
DX414499	<0.001	<0.001	Teosinte	NA	NA
DX414502	<0.001	0.024	Teosinte	0.3318	—
DX414461	NA	NA	NA	0.2566	—
DX414503	0.099	0.736	—	0.3036	—
DX414463	0.038	0.984	Domestication	0.2194	—
DX414535	0.028	0.005	Teosinte	0.7324	—
DX414464	NA	NA	NA	0.0606	—
DX414504	<0.001	0.025	Teosinte	0.9824	—
DX414465	<0.001	<0.001	Teosinte	0.7458	—
DX414505	0.003	0.032	Teosinte	0.928	—

(*continued*)

TABLE 3

(Continued)

	P-values in HKA test			*P*-values in CS test	
Gene	Maize	Teosinte	Selection status	Maize *vs.* teosinte	Selection status
DX414466	NA	NA	NA	0.3224	—
DX414467	<0.001	0.001	Teosinte	0.3868	—
DX414468	NA	NA	NA	0.9518	—
DX414469	0.666	0.943	—	0.4188	—
DX414506	NA	NA	NA	0.0026	Domestication
DX414507	NA	NA	NA	0.1228	—
DX414536	<0.001	0.496	Domestication	0.0418	Domestication
DX414508	0.998	0.738	—	0.6208	—
DX414537	<0.001	0.714	Domestication	0.0116	Domestication
DX414538	<0.001	0.585	Domestication	0.259	—
DX414539	NA	NA	NA	NA	NA
DX414472	<0.001	0.902	Domestication	0.04	Domestication
DX414540	NA	NA	NA	0.0734	—
DX414477	NA	NA	NA	0.8482	—
DX414541	<0.001	0.002	Teosinte	0.9666	—
DX414509	NA	NA	NA	0.0636	—

NA indicates that an outgroup was not available or that the test was not applicable due to the lack of segregating sites in teosinte. A dash (—) denotes that there was no evidence of selection in either maize or teosinte. Domestication genes and teosinte-selected genes are designated as domestication and teosinte, respectively.

To assess whether regulatory genes were a major target of selection during domestication, we compared the proportion of domestication genes identified in the unfiltered candidate gene pool to that found in our regulatory genes. Although the proportion of putative domestication genes identified was slightly higher in our regulatory candidate gene pool (32.6%) compared to the unfiltered candidates (17.1%), this difference was not significant (Fisher's exact test, $P = 0.19$). Similarly, there was no significant difference (Fisher's exact test, $P = 0.48$) between the number of putative domestication genes identified by the CS test in our study (14.7%) as compared to the number identified in the previous study (17.5%). These results provided no evidence that regulatory genes were a more frequent target of selection during domestication.

We used a similar procedure to assess whether regulatory genes were a major target of selection in teosinte. The HKA test identified statistically similar proportions of teosinte-selected genes in both our regulatory candidate gene pool (48.8%) and the unfiltered candidate genes (31.4%) (Fisher's exact test, $P = 0.16$). From this analysis, we found no convincing evidence supporting regulatory genes as a more frequent target of natural selection in teosinte.

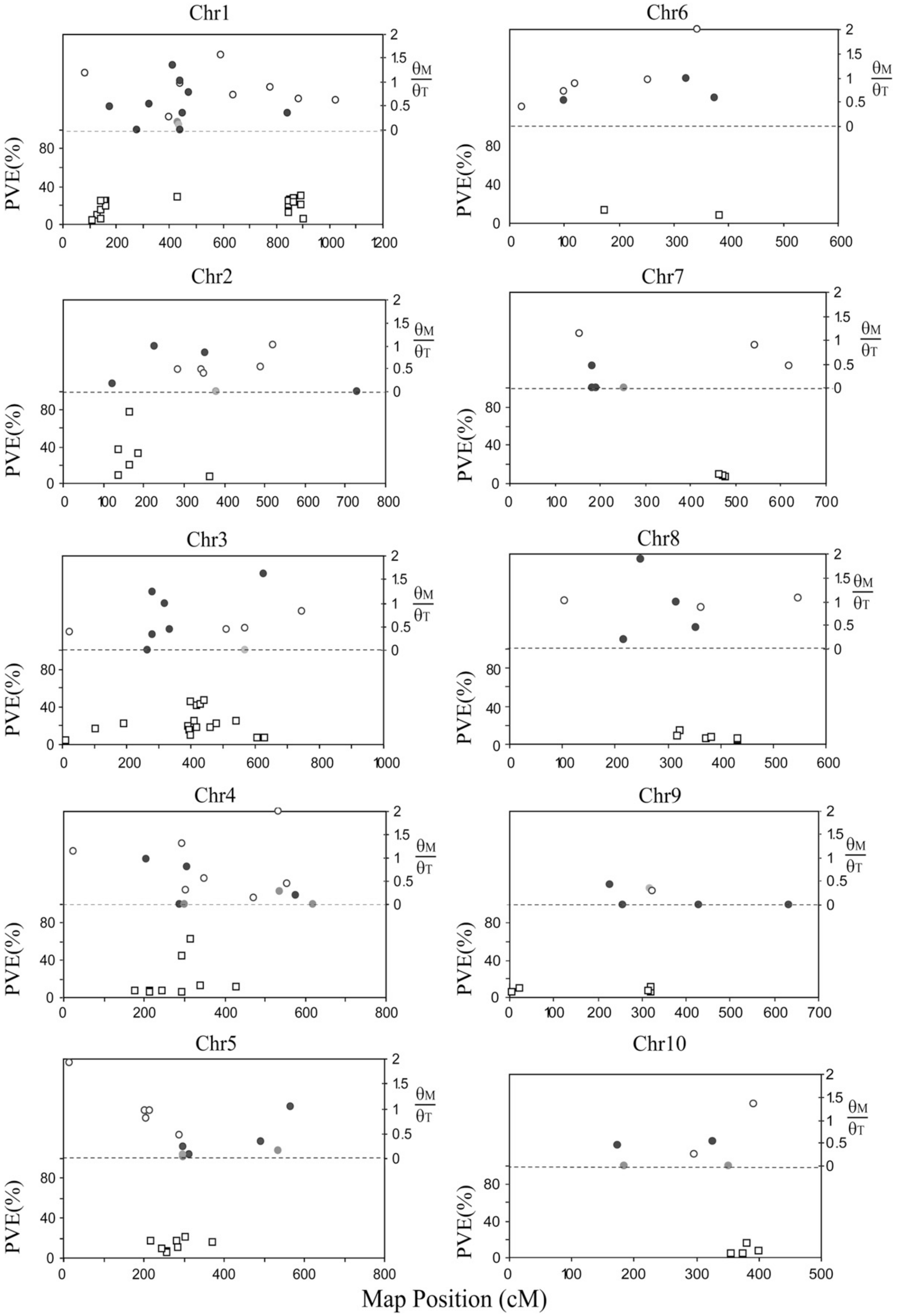
Finally, given that we have relatively few selected genes and thus low power to detect any differences in these proportions, we pooled genes identified by the HKA test as selected in either maize or teosinte into one group. We then compared the proportion of pooled selected genes among the regulatory candidates (81.4%) to that among the unfiltered candidates (48.6%) and observed a significant (Fisher's exact test, $P = 0.0035$) excess of selected genes among the regulatory candidate genes. This result provides some weak evidence that regulatory genes are overrepresented among selected genes and suggests that the negative results described above are due to low power. We consider the evidence weak

TABLE 4

Comparison of sequence statistics between 35 unfiltered candidate and 72 regulatory candidate genes

	Unfiltered candidate genes				Regulatory candidate genes			
	N^a	L^b	H^c	π^d	N^a	L^b	H^c	π^d
Maize landraces	14.3	291	2.2	0.0017	12.8	540	2.4	0.0014
Teosinte	14.8	297	4.9	0.0048	12.5	540	5.1	0.0042

[a] Average no. of sequences in the alignment.
[b] Average length of alignments, excluding gaps.
[c] Average no. of haplotypes in the alignments.
[d] Average amount of nucleotide diversity.


Chr1
Chr6
Chr2
Chr7
Chr3
Chr8
Chr4
Chr9
Chr5
Chr10
PVE(%)
$\frac{\theta_M}{\theta_T}$
0
40
80
0.5
1
1.5
2
Map Position (cM)

because our study may have had more power than Yamasaki *et al.* (2005) to detect selected genes due to the fact that our alignments are twice as long as those analyzed in their study.

Association of selected genes with maize domestication QTL: To assess if the domestication genes identified in our analysis were potentially causative for previously identified domestication QTL controlling morphological traits, we tested for association between the map positions of the two groups. Genetic map positions were obtained for 43 control and 60 candidate genes, which included 14 genes identified as targets of selection during domestication by the HKA test and/or the CS test (Figure 1). We used a permutation test to assess if the 14 putative domestication genes as a whole were more closely located to domestication QTL than a random sample of genes (Wright *et al.* 2005). The *P*-value of the permutation test ($P = 0.15$) was calculated as the proportion of the 100,000 random samples whose mean distances were as small or smaller than that observed for the domestication genes. This test provided no evidence for a significant clustering of putative domestication genes near known domestication QTL controlling morphological traits.

Among the 14 mapped domestication genes, 6 encode a kinase or phosphatase, which are often components in signal transduction pathways. Again, we used a permutation test to assess whether this subset of domestication genes significantly cluster near known domestication QTL. The average distance between these 6 kinase/phosphatase genes to the nearest domestication QTL is significantly smaller than the same distance calculated using a random sample of genes ($P = 0.006$). To verify that this phenomenon was not exclusively due to the function of these genes, we conducted the permutation test using 10 neutral kinase/phosphatase genes. We found that the neutral kinase/phosphatase genes did not significantly cluster near domestication QTL ($P = 0.72$). This suggests that the observed correlation between selected kinase/phosphatase genes and domestication QTL locations is not merely due to the function of these genes. In addition, we directly compared the average distances of the 6 selected and 10 neutral kinase/phosphatase genes to the closest domestication QTL and observed a significant difference ($P = 0.029$). These results suggested that the class of kinase/phosphatase genes under selection during domestication may be important contributors to the morphological divergence between maize and teosinte.

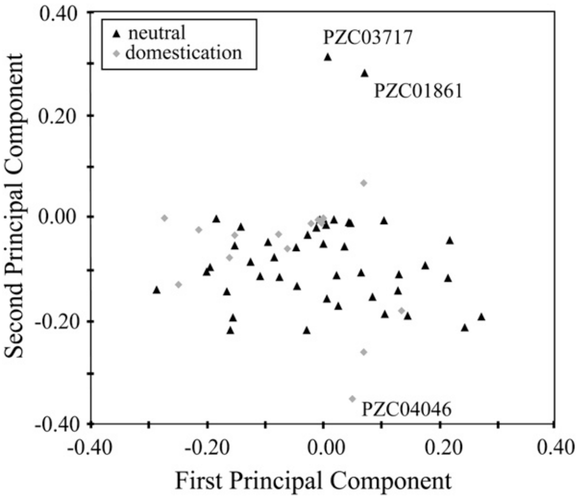


Figure 2.—Plot of the two leading principal components describing gene expression patterns in 17 maize tissues. Calculation was done using a covariance matrix. Solid triangles, neutral genes; shaded diamonds, domestication genes. Three outliers are labeled in the plot.

Expression analysis: Expression profiles of 66 genes in 17 distinct maize tissues were obtained from a MPSS database (see materials and methods). This set of 66 genes consisted of 16 putative candidate domestication genes, 23 neutral candidate genes, and 27 neutral control genes. As an initial form of comparison, we used PCA to detect any patterns in the expression profiles of domestication and neutral genes. The two leading principal components collectively explained 63.4% of the overall variation in expression. When the two leading principal components were plotted, domestication genes were distributed with neutral genes in a common cluster (Figure 2). By plotting the two leading principal components, three outliers, two neutral genes, and one putative domestication gene were identified. The two neutral genes (PZC03717 and PZC01861) are expressed at high levels in pollen as opposed to other tissues while the domestication gene (PZC04046) is expressed at high levels in all tissues except pollen. These results suggest that domestication genes do not have an expression profile across different tissues that is distinct from neutral genes due to the fact that they have experienced different selection histories during domestication.

A second set of analyses was conducted to determine the effects of gene class and tissue type on the level of gene transcript abundance. The 17 distinct tissues were grouped into three major tissue types: vegetative, ker-

Figure 1.—Map positions of 103 mapped genes analyzed in this article and QTL responsible for morphological divergence between maize and teosinte along the 10 maize chromosomes. The *x*-axis denotes chromosome position. The left *y*-axis denotes the percentage of phenotypic variance explained by a QTL and the right *y*-axis the ratio of nucleotide polymorphism in maize compared to that in teosinte. A horizontal (dashed) line is placed where the ratio of nucleotide polymorphism is 0. Open squares, chromosome locations of QTL; shaded circles, neutral genes; solid circles, domestication genes encoding a kinase/phosphatase; open circles, domestication genes whose protein products do not show sequence homology to a kinase/phosphatase.

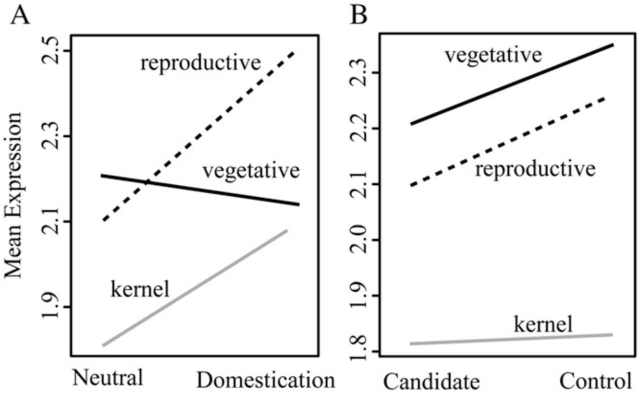


Figure 3.—Effects of gene class and tissue type on gene expression levels. The *y*-axis denotes the mean expression and is a logarithmic scale. Reproductive refers to reproductive tissues excluding kernels. (A) Mean expression level of 16 domestication candidate genes and 23 neutral candidate genes in three tissue types. (B) Mean expression level of 23 neutral candidate genes and 27 neutral control genes in three tissue types.

nel, and nonkernel reproductive tissues. Outliers in the PCA plot (Figure 2) were excluded from analysis due to the concern that they may skew the results. Permutation *t*-tests were conducted to determine if there was any significant difference between the expression of domestication and neutral genes in the three tissue types. Neutral and domestication genes are expressed at similar levels ($P = 0.61$) in vegetative tissues; however, domestication genes are expressed at a significantly (or near significantly) higher level than neutral genes ($P = 0.010$, $P = 0.069$) in kernel and other reproductive tissues (Figure 3A). To directly compare the relationship between neutral and domestication genes in vegetative tissue to that observed in kernel and other reproductive tissues, we performed ANOVA. There is a significant or nearly significant interaction between tissue type and gene class when comparing vegetative tissue to either kernel or other reproductive tissue (vegetative *vs.* kernel, $P = 0.053$; vegetative *vs.* reproductive tissues excluding the kernel, $P = 0.003$) (Figure 3A; Table 5). This illustrates that on average the expression of domestication genes is higher than that of neutral genes in both kernel and other reproductive tissues but not in vegetative tissues.

We conducted tests similar to those above to determine if gene classification (control *vs.* candidate) was associated with the level of gene expression in different tissue types. Control genes have a slightly but not significantly higher level of expression than candidate genes in all three tissue types (permutation *t*-test: vegetative, $P = 0.21$; kernel, $P = 0.90$; reproductive tissues excluding the kernel, $P = 0.26$; Figure 3B). These results, as well as the ANOVA, suggest that there is no significant difference in expression of control and candidate genes in the three tissue types (Table 5).

TABLE 5

Effect of interaction between gene class and tissue type on gene expression levels

	Gene class			
	Neutral *vs.* domestication genes		Candidate *vs.* control genes	
Tissue type	χ^{2a}	*P*-value	χ^{2a}	*P*-value
Vegetative *vs.* reproductive[b]	8.9	0.003**	2.2	0.138
Vegetative *vs.* kernel	3.7	0.053	1.4	0.237
Reproductive[b] *vs.* kernel	1.2	0.273	1.0	0.317

Gene class has two levels: neutral or selected. The *P*-value was evaluated from ANOVA results by fitting a linear mixed-effects model where gene class and tissue type were considered as fixed effects while individual genes and libraries were considered random effects. Significant *P*-values are designated as $*P < 0.05$, $**P < 0.01$, and $***P < 0.001$.

[a] The χ^2 value was computed as the log-likelihood ratio of model 1 and model 2 (see materials and methods).

[b] "Reproductive" refers to reproductive tissues excluding kernels.

DISCUSSION

In this study, we tested 72 regulatory genes for signatures of selection. Among these 72 genes, we identified 17 putative domestication genes as well as 21 genes with evidence of selection in teosinte. To determine if regulatory genes as a whole were major targets of selection, we compared the results of our study with a similar study where the candidate genes had not been filtered in regard to function. From this comparison, we found that regulatory genes were significantly enriched for selected genes; however, there may be a difference in power between the two studies. To assess the specific role of regulatory genes during domestication, we investigated whether these genes were located near previously identified domestication QTL controlling morphological change. Although no significant association was found between the genetic map positions of regulatory domestication genes and domestication QTL, a subset of selected regulatory genes, kinases and phosphatases, did colocalize with domestication QTL. This result suggests that kinases and phosphatases may contribute to the morphological divergence between maize and teosinte. Finally, we found that putative domestication genes as compared to neutral genes were expressed at higher levels in reproductive tissue.

Search for targets of selection: We used both the HKA test and the CS test to identify domestication genes within our candidate pool. Fourteen putative domestication genes were identified by the HKA test and 10 putative domestication genes were identified by the CS test. Seven of these genes were identified by both tests, providing more evidence that they underwent selection during domestication. In total, these two tests identified

17 putative domestication genes. The HKA test was also used to identify 21 genes that underwent selection in teosinte. Similar to that observed in a previous study (Yamasaki *et al.* 2005), our strategy of examining genes with low diversity in maize inbred lines resulted in a much larger proportion of identified putative domestication genes (∼23.6%) than that reported by previous studies (∼2.0%, Vigouroux *et al.* 2002; ∼2–4%, Wright *et al.* 2005) that examined genes at random.

Although our strategy of identifying putative selected genes among candidate genes with low diversity in maize was successful, our results should be considered with caution. The tests that we used to identify putative selected genes are tests of neutrality; thus a significant result simply implies deviation from neutrality, which is not necessarily due to selection. Population structure and history can also lead to a deviation from neutrality. The CS test takes one aspect of the history of the population (the domestication bottleneck) into account; however, the results rely on modeling this aspect accurately and ignore other aspects of population history. The HKA test also compensates for the domestication bottleneck since the control genes have experienced the same history as the test gene. Another issue is that our selected genes may be hitchhiked regions of selective sweeps on neighboring genes rather than direct targets of selection, and our approach does not have power to distinguish between the two possibilities. Thus, we emphasize that the selected genes identified through these analyses are candidates and that further analyses are needed to validate that these genes were targets of selection.

The role of regulatory genes in evolution: One goal of this study was to test the hypothesis that regulatory genes are overrepresented among selected genes. Our results provide limited support for this hypothesis. A direct comparison found that our study identified a significantly larger proportion of selected genes as compared to a previous study using a similar strategy with candidate genes that were unfiltered in regard to function (Yamasaki *et al.* 2005); however, this difference could be due to a difference in the power to detect selected genes between the two studies. An overrepresentation of regulatory genes among selected genes is not surprising, given the number of selected regulatory genes that control domestication traits (Doebley *et al.* 1997; Frary *et al.* 2000; Wang *et al.* 2005; Li *et al.* 2006; Simons *et al.* 2006). Regulatory genes are obvious candidates for controlling traits that have undergone selection due to the fact that changes in function or expression of regulatory genes can potentially change the expression of downstream structural or regulatory genes. Further validation of the putative regulatory selected genes in this study as well as the identification of more putative selected genes in maize and other organisms will be necessary to determine if regulatory genes are a more frequent target of selection in general.

Domestication genes and maize–teosinte QTL: Some of our 17 candidate domestication genes colocalize with QTL controlling maize–teosinte morphological divergence (Figure 1). Although we observed that the average distance to the closest domestication QTL is smaller for domestication genes than for neutral genes, this difference is not statistically significant. We consider several reasons that may explain the weak statistical evidence. First, the method we used for comparison may not be powerful enough, given that we simply consider the distance to the nearest QTL without taking into account clustered QTL and the size of the QTL effect. Second, QTL used here explain only the variation of visible morphological traits (Doebley and Stec 1991, 1993) and are not expected to be associated with domestication genes selected for other traits. Third, some domestication genes might be QTL with minor effects or QTL not expressed stably across environment. These types of QTL are not easily detected with certainty and may not have been identified in earlier studies (Doebley and Stec 1991, 1993).

We did observe a significant association between the map positions of selected kinase/phosphatase genes with those of known domestication QTL. We hypothesize that kinase/phosphatase genes may have been targets of selection for morphological change during domestication due to the importance of such genes in regulating plant development through signal transduction pathways. Although a kinase has been found to be responsible for flowering-time differences among varieties in rice (Takahashi *et al.* 2001), to date, no crop domestication QTL has been fine mapped to a kinase or phosphatase. This class of genes represents a good candidate class that should be considered in future candidate gene analyses.

Gene expression patterns: One of the intriguing questions in evolutionary biology is how variation in gene expression patterns contributes to evolution. During the evolution of maize, artificial selection acted on some genes, which regulate divergent traits between maize and teosinte, while other genes evolved under neutrality without contributing to domestication. Genes under selection are expected to be expressed in tissues that differ in morphology between maize and teosinte, while neutral genes could be expressed in any tissue type. This hypothesis is also supported by a recent study that examined expression levels of 48 selected genes and 658 neutral genes and concluded that selected genes were more highly expressed in the ear, a tissue that is strikingly different in maize and teosinte (Hufford *et al.* 2007). In our study, we observed a similar phenomenon: substantially stronger expression for selected genes as opposed to neutral genes in the kernel and other reproductive tissues, but not in vegetative tissues.

The results from this study have four important implications that should be considered in the search for selected genes. First, the comparison of several studies have indicated that the strategy of choosing candidate genes with low diversity in maize inbreds has resulted in a much higher proportion of identified putative selected genes. Second, we observed that regulatory genes are overrepresented among selected genes. Our results also suggest that a subset of regulatory genes, kinases and phosphatases, may have been targets of selection for morphological change during domestication. Third, our analysis of the expression profiles of domestication and neutral genes within reproductive tissues supports the inference that selected genes are expressed at a higher level in these tissues as compared to neutral genes. This result suggests that a substantial portion of the domestication genes identified by our study are real as opposed to false positives. Finally, although at present the number of putative domestication genes is small, the approaches used in this study and those described elsewhere, as well as advancements in sequencing technology, will make the identification of domestication genes easier in the near future. As the list of identified domestication genes grows, so will our understanding of the underlying process of domestication.

This work was funded by National Science Foundation grant DBI-0321467, National Institutes of Health grant GM-58816, and U. S. Department of Agriculture Hatch grant WIS04772.

LITERATURE CITED

Bates, D., 2007 Linear mixed-effect models using S4 classes. R Package, Version 0.99875-7. http://cran.r-project.org/web/packages/lme4/index.html

Brenner, S., M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd *et al.*, 2000 Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat. Biotechnol. **18:** 630–634.

Clark, R. M., E. Linton, J. Messing and J. F. Doebley, 2004 Pattern of diversity in the genomic region near the maize domestication gene *tb1*. Proc. Natl. Acad. Sci. USA **101:** 700–707.

Doebley, J., and L. Lukens, 1998 Transcriptional regulators and the evolution of plant form. Plant Cell **10:** 1075–1082.

Doebley, J., and A. Stec, 1991 Genetic analysis of the morphological differences between maize and teosinte. Genetics **129:** 285–295.

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., A. Stec and L. Hubbard, 1997 The evolution of apical dominance in maize. Nature **386:** 485–488.

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

Frary, A., T. C. Nesbitt, S. Grandillo, E. Knaap, B. Cong *et al.*, 2000 *fw2.2*: a quantitative trait locus key to the evolution of tomato fruit size. Science **289:** 85–88.

Hall, T., 1999 BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids. Symp. Ser. **41:** 95–98.

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

Hudson, R. R., 1987 Estimating the recombination parameter of a finite population model without selection. Genet. Res. **50:** 245–250.

Hudson, R. R., M. Kreitman and M. Aguadé, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

Hufford, K. M., P. Canaran, D. H. Ware, M. D. McMullen and B. S. Gaut, 2007 Patterns of selection and tissue-specific expression among maize domestication and crop improvement loci. Plant Physiol. **144:** 1642–1653.

Li, C., A. Zhou and T. Sang, 2006 Rice domestication by reducing shattering. Science **311:** 1936–1939.

Purugganan, M. D., 1998 The molecular evolution of development. BioEssays **20:** 700–711.

Purugganan, M. D., 2000 The molecular population genetics of regulatory genes. Mol. Ecol. **9:** 145–1461.

Rozas, J., J. C. Sanchez-DelBarrio, X. Messeguer and R. Rozas, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:** 2496–2497.

Simons, K. J., J. P. Fellers, H. N. Trick, Z. Zhang, Y. S. Tai *et al.*, 2006 Molecular characterization of the major wheat domestication gene *Q*. Genetics **172:** 547–555.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

Takahashi, Y., A. Shomura, T. Sasaki and M. Yano, 2001 *Hd6*, a rice quantitative trait locus involved in photoperiod sensitivity, encodes the alpha subunit of protein kinase CK2. Proc. Natl. Acad. Sci. USA **98:** 7922–7927.

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

Wang, H., T. Nussbaum-Wagler, B. Li, Q. Zhao, Y. Vigouroux *et al.*, 2005 The origin of the naked grains of maize. Nature **436:** 714–719.

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., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Wright, S. I., and B. S. Gaut, 2005 Molecular population genetics and the search for adaptive evolution in plants. Mol. Biol. Evol. **22:** 506–519.

Wright, S. I., I. V. Bi, S. G. Schroeder, M. Yamasaki, J. F. Doebley *et al.*, 2005 The effects of artificial selection on the maize genome. Science **308:** 1310–1314.

Yamasaki, M., M. I. Tenaillon, I. V. Bi, S. G. Schroeder, H. Sanchez-Villeda *et al.*, 2005 A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. Plant Cell **17:** 2859–2872.

Communicating editor: J. A. Birchler