Estimating a Nucleotide Substitution Rate for Maize from Polymorphism at a Major Domestication Locus

Richard M. Clark,[*1] Simon Tavaré,†‡ and John Doebley**

*Laboratory of Genetics, University of Wisconsin; †Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California; and ‡Department of Oncology, University of Cambridge, Cambridge, United Kingdom

To estimate a rate for single nucleotide substitutions for maize (*Zea mays* ssp. *mays*), we have taken advantage of data from genetic and archaeological studies of the domestication of maize from its wild ancestor, teosinte (*Z. mays* ssp. *parviglumis*). Genetic studies have shown that the *teosinte branched1* (*tb1*) gene was a major target of human selection during maize domestication, and sequence diversity in the intergenic region 5′ to the *tb1*-coding sequence is extraordinarily low. We show that polymorphism in this region is consistent with new mutation following fixation for a small number of *tb1* haplotypes during domestication. Archeological studies suggest that maize was domesticated ~6,250–10,000 years ago and subsequently the size of the maize population is thought to have expanded rapidly. Using the observed number of mutations within the region of selection at *tb1*, the approximate age of maize domestication, and approximations for the maize genealogy, we have derived estimates for the nucleotide substitution rate for the *tb1* intergenic region. Using two approaches, one of which is a coalescent approach, we obtain rate estimates of $\sim 2.9 \times 10^{-8}$ and 3.3×10^{-8} substitutions per site per year. We also show that the pattern of polymorphism in the *tb1* intergenic region appears to have been strongly affected by the mutagenic effect of DNA methylation. Excluding target sites of symmetric DNA methylation (CG and CNG sites) from analysis, the mutation rate estimates are reduced by ~50%–60%, while the rates for CG and CNG sites are nearly an order of magnitude higher. We use rate estimates from the *tb1* region to estimate the timing of expansion of transposable elements in the maize genome and suggest that this expansion occurred primarily within the last million years.

Introduction

Mutation is the ultimate source of genetic variation, and the rate with which single DNA bases mutate (μ_{bp}) is an important parameter for population and evolutionary genetic studies. However, $\mu_{bp} < 1 \times 10^{-7}$ for most organisms (Drake et al. 1998; Kumar and Subramanian 2002; Nachman and Crowell 2000) and the low frequency of nucleotide substitution makes μ_{bp} difficult to estimate directly using mutation-accumulation designs. Consequently, most estimates of μ_{bp} are indirect and have been calculated using either of two approaches. In the first, the initial step is to calculate a per-locus mutation rate. The per–base pair mutation rate is then estimated by dividing the per-locus mutation rate by the number of bases in a gene that can be mutated to give an observable phenotype (see review by Drake et al. 1998). However, the latter quantity is almost never known. A further complication is that per-locus rates may also be influenced by other mutational processes such as changes in the length of simple sequence repeats or transposition of mobile elements. In some cases, the frequency of these events is known to be orders of magnitude higher than that of nucleotide substitution itself (e.g., Brinkmann et al. 1998; Vigouroux et al. 2002*a*).

The advent of DNA sequencing, along with the theoretical framework of the neutral theory of molecular evolution (Kimura 1983), has allowed a second and more widely applied approach to estimating the rate of nucleotide substitution. This "phylogenetic rate" approach uses the divergence time between lineages and the number of substitutions between orthologous sites to estimate μ_{bp}. This basic approach has been applied to many taxa (e.g., Gaut et al. 1996; Nachman and Crowell 2000) but has its own set of uncertainties. First, divergence times are estimated from the fossil record, which is often incomplete or ambiguous and is not available for some taxa. Second, rate estimates are often derived using species that have had differing generation lengths for at least a portion of their independent history. It has been widely hypothesized that in animals μ_{bp} should vary inversely with generation time because animals set aside germline cells whose replication number is largely independent of generation length. It should be noted, however, that the relationship between generation time and mutation rate in animals is still unclear (e.g., Kumar and Subramanian 2002). In plants, reproductive cells are not specified until flowering, and the relationship between generation time and mutation rate is even less well understood (Gaut et al. 1996). Third, the phylogenetic approach requires data from neutral sites, and polymorphism at synonymous sites is typically examined because synonymous sites can be aligned between distantly related sequences. Nevertheless, not all synonymous sites are neutral. The extent to which these factors bias rate estimates can be difficult to ascertain. Rate calculations based on recently diverged lineages with excellent fossil records can minimize some of these uncertainties (e.g., Nachman and Crowell 2000).

Cross-disciplinary investigations of plant domestication by humans (Smith 2001) provide a complementary framework to traditional approaches to ask questions about mutation rates. A major finding of quantitative genetic studies examining the domestication process has been that a small number of loci can explain much of the difference between domesticates and their wild ancestors in several crop species (Doebley and Stec 1991, 1993; Doganlar et al. 2002). If alleles at major effect loci were fixed rapidly by human selection during domestication, then extant diversity at these loci should be due to new mutation subsequent to domestication. In general, crop species are thought to have undergone rapid population expansion following

[1] Present address: Max-Planck-Institute for Developmental Biology, Department VI—Molecular Biology, Tübingen, Germany.

Key words: nucleotide, substitution, mutation, maize, methylation, coalescent.

E-mail: richard.clark@tuebingen.mpg.de.

Mol. Biol. Evol. 22(11):2304–2312. 2005
doi:10.1093/molbev/msi228
Advance Access publication August 3, 2005

domestication, and the theoretical framework for the shape of genealogies following rapid expansion has been carefully quantified (Slatkin and Hudson 1991; Griffiths and Tavaré 1994). Given the number of mutations in a sample and the approximate time since domestication, the underlying mutation rate can be estimated. Domestication of crops has been recent enough that fairly precise estimates of domestication times are available from archeological studies (Smith 1998) but old enough that it should be possible to observe mutation accumulation with reasonably sized data sets.

The *tb1* gene of maize provides an opportunity for calculating μ_{bp} under this scenario. Maize was domesticated from its wild progenitor, teosinte, between ~6,250 and 10,000 years ago (Smith 1998; Piperno and Flannery 2001) in a single domestication event (Matsuoka et al. 2002). The *tb1* gene controls several aspects of plant architecture that differ between maize and teosinte (Doebley, Stec, and Hubbard 1997). A combination of quantitative trait locus-mapping studies, complementation tests, and molecular evolutionary studies has shown that *tb1* was a major target of human selection during domestication (Doebley and Stec 1993; Doebley, Stec, and Gustus 1995; Doebley, Stec, and Hubbard 1997; Wang et al. 1999). Recently, we showed that an intergenic region of 60–90 kb that extends 5′ from the *tb1* cDNA sequence appears to have been fixed for a small number of haplotypes during the domestication process (Clark et al. 2004). In the current study, we use polymorphism within the region of the selective sweep at *tb1* to derive several estimates of μ_{bp} for maize.

Materials and Methods

Sequence Data

To assess mutation number, we aligned maize sequences for six regions (1.7-, 2.5-, 7.1-, 35.6-, 45.8-, and 58.6-kb regions; distances are from the *tb1* cDNA sequence) within the core selective sweep at the *tb1* locus. Sample and sequence information for this data set has been documented by Clark et al. (2004), and we follow their naming conventions to allow direct comparison to the earlier study. All sequence data collected for the 2.5-, 7.1-, 35.6-, 45.8-, and 58.6-kb regions were generated by Clark et al. (2004) from direct sequencing (both strands) of polymerase chain reaction–amplified products. Sequences for the 1.7-kb site were previously determined by Tenaillon et al. (2001), and Clark et al. (2004) verified all singleton changes at the 1.7-kb region by resequencing from independently amplified products. The present study uses 23 of the 24 samples analyzed by Clark et al. (2004); one sample (Maize 15) was excluded from the present study because sequence for this sample had not been determined for the 58.6-kb region. In addition, sites included in a small number of short insertion/deletion polymorphisms, primarily associated with several simple sequence repeats, were excluded in calculating mutation rates (see below).

Classical Approach to Mutation Rate Estimation

Under a model of rapid population expansion, the underlying phylogeny is approximately star shaped (Slatkin and Hudson 1991) and each of n independent lineages has an approximately equal branch length. Let μ_{reg} be the mutation rate per generation for a chromosomal region of l base pairs. The number of mutations in g generations along a single lineage is the sum of g Bernoulli random variables, which has approximately a Poisson distribution with mean $\lambda = g\mu_{reg}$. The maximum likelihood estimator of λ is $\hat{\lambda} = \bar{M}$, where $\bar{M}$ is the average number of mutations observed per lineage, and the approximate variance of this estimator is given by $\bar{M}/n$. The estimator of μ_{reg} is $\hat{\mu}_{reg} = \bar{M}/g$, and the estimator of the per–base pair mutation rate μ_{bp} is $\hat{\mu}_{bp} = \hat{\mu}_{reg}/l$ with approximate variance $(\bar{M}/n)/(g^2 \times l^2)$.

A Bayesian Approach to Mutation Rate Estimation

We also adopt a coalescent-based approach to Bayesian inference about μ_{bp} similar to that of Tavaré et al. (1997). The underlying parameters are g, μ_{reg}, N_0 (the size of the ancestral population at the time of domestication), and N_1 (the size of the present day population). Time is measured in units of $2N_1$ generations, and we define the following scaled parameters:

1. $T = g/(2N_1)$, the time of domestication in coalescent units;
2. $\theta = 4N_1\mu_{reg}$, the rescaled mutation rate;
3. $\beta = T^{-1}\log(N_1/N_0)$, the population expansion rate assuming exponential growth from size N_0 to size N_1 in time T.

To simulate the ancestral history of a sample of n chromosomal regions, we proceed as follows (cf. Griffiths and Tavaré 1994). Let T_n, $T_{n\text{-}1}$, …, T_2 be independent exponential random variables, with expectation

$$ET_j = \bar{M} = \frac{2}{j(j-1)}, j = n, \ldots, 2.$$

Times are then generated according to the following scheme:

Set $J = n+1, S_J = S_J^v = 0$;
Step 1a: Set $J = J - 1$, $S_J = S_{J+1} + T_J$, and $S_J^v = \frac{1}{\beta}\log\{1+\beta S_J\}$;
Step 2a: If $S_J^v > T$, stop; else return to Step 1a.

The number of distinct ancestors time T ago is J, and the length of time for which there are j distinct ancestors in the sample is $T_j^v = S_j^v - S_{j+1}^v, j = n, n-1, \ldots, J + 1$. There are J ancestors for time $T_J^v = T - S_{J+1}^v$. The total length of the branches of the ancestral trees of the sample is then

$$L = \sum_{j=J}^{n} jT_j^v. \qquad (1)$$

Assuming an infinitely many sites mutation model, each mutation that occurs on the branches of these ancestral trees leaves a segregating site in the sample. The number of such sites is denoted by S. The following algorithm can be used to simulate observations from the posterior distribution of the parameter $\alpha = (g, \mu_{reg}, N_0, N_1)$, conditional

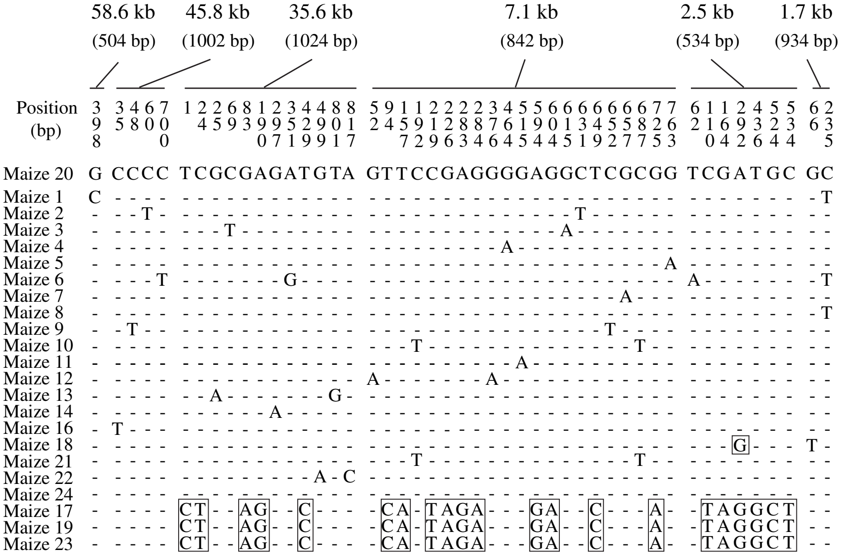

FIG. 1.—Polymorphism data for six sampled regions within the selective sweep at *tb1*. Region names (top) refer to distance in kilobases from the *tb1* cDNA sequence. Boxed nucleotides define a second haplotype present at the 2.5-, 7.1-, and 35.6-kb regions (see samples Maize 17, 19, and 23). The consensus sequence (Maize 20) is shown at the top.

on the number of segregating sites $S = s$ observed in the sample.

Step 1b: Generate values for α from its prior distribution;

Step 2b: Calculate T, θ, and, β as above;

Step 3b: Simulate the coalescent history of the sample back to time T (as described above) using these parameter values;

Step 4b: Calculate the total branch length L according to equation (1); and

Step 5b: Calculate:

$$h = \frac{\mathrm{Po}(L\theta/2)\{s\}}{\mathrm{Po}(s)\{s\}};$$

Step 6b: Accept α with probability h; go to Step 1b.

In Step 5b, the term $\mathrm{Po}(\lambda)\{s\}$ denotes the Poisson probability

$$\mathrm{Po}(\lambda)\{s\} = \mathrm{e}^{-\lambda}\lambda^s/s!.$$

The accepted values produced by this algorithm form a random sample from the required posterior.

All estimates are based on 50,000 accepted observations, and values of μ_{reg} used for simulation were selected uniformly such that they corresponded to μ_{bp} ranging from 10^{-10} to 10^{-6} (note that $\mu_{reg} = \mu_{bp} \times l$). This range includes virtually all estimates of μ_{bp} from animals and plants (e.g., Gaut et al. 1996; Nachman and Crowell 2000; Kumar and Subramanian 2002). Investigation of the maize domestication event has suggested that the founding size of the maize population may have been between several tens to several thousands of individuals depending on the demographic scenario (Eyre-Walker et al. 1998), and for the present study we selected N_0 uniformly between 20 and 10,000. We selected N_1 uniformly between 10^6 and 10^9, consistent with the large extant population size and rapid expansion following initial domestication. The distribution of g used for simulations is discussed in *Results*.

Results

The core of the selective sweep at *tb1* extends 60–90 kb 5′ to the *tb1* gene and is entirely contained within the 161-kb intergenic region between *tb1* and the nearest upstream gene (Clark et al. 2004). An alignment of maize sequences for six regions of approximately 500–1,000 bp each that are located within the core selective sweep is shown in figure 1 (region names are given in distance in kilobases from the *tb1* cDNA sequence). Each of these regions contains largely unique sequence, and all sites within these regions are presumed to be silent (Clark et al. 2004). Across all regions, 4,840 sites are represented (table 1), and a single major haplotype is present at the 1.7-, 45.8-, and 58.6-kb regions. For the 2.5-, 7.1-, and 35.6-kb regions, 20 samples belong to a major haplotype, whereas 21 nucleotide differences define a second distinct haplotype represented by samples 17, 19, and 23 (fig. 1; nucleotide changes distinguishing major haplotypes are open boxed). Therefore, across the core selective sweep, either one or two major haplotypes appear to have come through the domestication event (see also Clark et al. 2004). The 21 nucleotide differences that distinguish the major haplotypes are presumed to have arisen prior to maize domestication and were excluded from subsequent analyses.

Number and Pattern of Mutations

Across all regions, 26 mutations are observed (table 1 and fig. 1), and the ratio of transitions to transversions is 4.2. In general, intergenic maize regions have high levels of DNA methylation (Palmer et al. 2003). Methylation in plants occurs at C residues (Finnegan et al. 1998), and deamination of 5-methyl deoxycytidine to thymidine is

Table 1
Mutation Summary

Region (kb)	Sites (bp)	Mutations	Transversions	Transitions	Transitions at CG or CNG Sites	Total CG and CNG Sites
1.7	934	2	1	1	1	94
2.5	534	1	1	0	0	98
7.1	842	11	0	11	11	191
35.6	1,024	7	2	5	4	112
45.8	1,002	4	0	4	0	53
58.6	504	1	1	0	0	122
All regions	4,840	26	5	21	16	670

thought to lead to transitions at a high frequency (Coulondre et al. 1978). To examine whether this mutational process may have affected the pattern of mutation in the *tb1* intergenic region, we determined the number of transitions that occurred at target sites of symmetric C methylation (CG and CNG) (Finnegan et al. 1998). Of the 26 mutations in the *tb1* intergenic region, 62% are transitions at CG or CNG sites, even though these sites account for only 14% of all sites (table 1). This pattern is particularly striking for the 7.1- and 35.6-kb regions for which 15 of 18 mutations are transitions at CG or CNG sites. Across all regions, the elevated frequency of mutations at CG and CNG sites is highly significant ($\chi^2 = 49.8$, df = 1, $P < 0.0001$), consistent with an effect of methylation on the mutational process in the *tb1* region. If transitions at CG and CNG sites are excluded, a 1:1 transition to transversion ratio is observed across all regions. No transversions occurred at CG or CNG locations.

The observed number of mutations at each of the six regions differs marginally from the values expected if mutations are distributed randomly among regions ($\chi^2 = 14.1$, df = 5, $P < 0.05$; expected values were calculated by weighting by the number of sites per region) and is indicative of a difference in mutation rates among regions. Because of the strong bias for transitions at CG and CNG sites (see above), and variation for CG/CNG content among regions (table 1), we performed an analogous test where the expected number of mutations per region was weighted by both the length of the region and the CG/CNG content. This test was not significant. However, the *P* value remained low ($\chi^2 = 9.9$, df = 5, $P = 0.08$), and it is possible that differential representation of CG and CNG sites between regions may partly account for the observed nonrandom distribution of mutations between regions.

New Versus Preexisting Mutations

An assumption for our approach to estimating mutation rates using polymorphism data from the *tb1* region in maize is that we have correctly identified mutations that have arisen subsequent to fixation of *tb1* haplotypes during the initial domestication event. If substitutions in the *tb1* intergenic region arose by new mutation following domestication, they should be absent (or at least underrepresented) among teosinte samples. We tested this for the 17 mutations at the 1.7-, 7.1-, and 45.8-kb regions for which Clark et al. (2004) also sampled sequence diversity from ssp. *parviglumis*, the presumed progenitor to maize (Matsuoka et al. 2002). One mutation at the 45.8-kb region was represented in ssp. *parviglumis*, and 3 of the 11 mutations at the 7.1-kb region were present in ssp. *parviglumis* sequences.

To assess whether these mutations are likely to have been introduced into the maize data set from standing variation that predates the maize domestication event, we examined the structure of the teosinte haplotypes that harbor polymorphisms in common with maize. The teosinte haplotypes harboring shared polymorphisms are different from the corresponding maize sequences (4–13 nucleotide or indel differences per region, unpublished data). This observation argues against introduction of these shared polymorphisms into the maize sample as a result of selection for closely related haplotypes that also share the functional site or sites at *tb1* selected during maize domestication. In addition, for each shared polymorphism, the teosinte haplotype has an additional nucleotide change (or changes) within 5–94 bp of the shared site that is not present in the corresponding maize sample (unpublished data). In turn, the absence of these nearby polymorphisms in the maize sequences argues against the introduction of nucleotide changes into maize by either recombination or gene conversion from unselected haplotypes.

We believe therefore that the most parsimonious explanation for the presence of shared polymorphisms in the data set of Clark et al. (2004) is new mutation in the maize lineage at sites polymorphic in teosinte. This is particularly plausible for the three shared polymorphisms at the 7.1-kb region, where 6% of all CG and CNG sites were mutated in the maize sample alone, and diversity in teosinte is much higher than in maize (Clark et al. 2004) (but see the caveats in *Discussion*).

Estimating μ_{bp} for Maize

We have used two approaches to estimate μ_{bp} from intergenic *tb1* sequences assuming rapid expansion of the maize population following domestication. In the first, we assume a star phylogeny in which all sampled lineages coalesce at the time of fixation of *tb1* haplotypes during domestication. Using this model, μ_{bp} can be estimated from the observed number of mutations and an estimate of the time since fixation (see *Materials and Methods*). We refer to this as the "classical approach." A small number of haplotypes were sampled repeatedly in our data set (e.g., samples Maize 10 and 21, fig. 1), and therefore the assumption of a star phylogeny holds only approximately (i.e., going back in time, some lineages coalesce before the fixation event). In this case, the total length of the underlying tree (the sum of the lengths of all branches) is reduced. Because

Table 2
Mutation Rate Estimates for *tbl* Intergenic Regions

			Mutation Rate (μ_{bp}) Estimation ($\times 10^{-8}$)			
			Classical Approach[a]	Bayesian Approach[c]		
Region (kb)	Length (bp)	Mutation Number	$\hat{\mu}_{bp}$ ± 1.96 s.d.[b]	LCL (95%)	$\hat{\mu}_{bp}$	UCL (95%)
1.7	934	2	1.2 ± 1.6	0.4	1.9	4.8
2.5	534	1	1.0 ± 2.0	0.3	2.2	6.3
7.1	842	11	7.2 ± 4.3	4.2	8.5	14.8
35.6	1,024	7	3.7 ± 2.8	1.9	4.7	8.8
45.8	1,002	4	2.2 ± 2.1	0.9	3.0	6.3
58.6	504	1	1.1 ± 2.1	0.3	2.4	6.7

[a] Mutation rate estimates for the classical approach are for $g = 8{,}000$.

[b] s.d., standard deviation; LCL, lower credible level; UCL, upper credible level.

[c] Values of $\hat{\mu}_{bp}$ for the Bayesian approach are the mean of 50,000 accepted observations and error estimates are the 95% LCL and UCL.

the number of mutations is unchanged, but the total length of the phylogeny is reduced, the classical estimate of μ_{bp} is likely to underestimate the actual substitution rate. To correct for repeated sampling of the haplotype or haplotypes that share a part of their histories, we have also estimated μ_{bp} using a Bayesian approach based on a coalescent framework (hereafter referred to as the "Bayesian approach") that does not assume a star-shaped phylogeny (see *Materials and Methods*).

Both approaches require an estimate for the fixation time of haplotypes within the core selective sweep at *tb1*, and this can be approximated by the time of maize domestication. The earliest archaeological evidence of maize is 6,250 years ago (Piperno and Flannery 2001), and the earliest evidence of domestication of any crop in the western hemisphere is 10,000 years ago (Smith 1998). Unless otherwise noted, for the classical approach, we assume that a reasonable estimate of the age of domestication is 8,000 years (which also corresponds to ~8,000 generations because *Zea mays* is an annual). For the Bayesian approach, generation number (g) was selected uniformly from 6,250 to 10,000 for coalescent simulations. Estimates of μ_{bp} for the six regions using the classical approach range from 1.0×10^{-8} to 7.2×10^{-8}, and confidence intervals (CIs) show substantial overlap (table 2). Applying the Bayesian approach, the means of the posterior estimates for μ_{bp} range from 1.9×10^{-8} to 8.5×10^{-8} (table 2), and 95% credible intervals for the smallest and largest estimates overlap.

We also calculated three rate estimates using data combined from all six regions. For the first estimate, all mutations and all sites were included to give a "general rate" (table 3, top). Second, because of the significant bias for mutations at CG and CNG sites (table 1), we calculated a mutation rate for target sites of symmetric methylation ("methylated rate"; table 3, middle). Because we do not know what frequency of CG and CNG sites are actually methylated in the *tb1* intergenic region, this rate should be interpreted with caution. Finally, we calculated the rate for sites that are not predicted targets of symmetric methylation ("nonmethylated rate"; table 3, bottom). The general, methylated, and nonmethylated rates as a function of generation number from 6,250 to 10,000 are also shown based on the classical approach (fig. 2). Additionally, for the Bayesian approach, the distribution of the posterior values for $\hat{\mu}_{bp}$ are shown for the general rate estimation (fig. 3). Despite the wide and uniform prior distributions used for the Bayesian approach, values for $\hat{\mu}_{bp}$ are tightly clustered.

While the classical and Bayesian estimates are based on different approaches, differences between the corresponding estimates are small (<19% for calculations using all regions), and a given estimate is typically included within the 95% credible interval for the corresponding estimate (tables 2 and 3). Because the Bayesian approach does not require simultaneous coalescence of all lineages, we believe that the Bayesian estimates, whose absolute values are marginally higher than those estimated using the classical approach, may more accurately reflect underlying mutation rates.

Table 3
Combined Rate Estimates

			Mutation Rate (μ_{bp}) Estimation ($\times 10^{-8}$)			
			Classical Approach[a]	Bayesian Approach[c]		
Sites Analyzed	Length (bp)	Mutation Number	$\hat{\mu}_{bp}$ ± 1.96 s.d.[b]	LCL (95%)	$\hat{\mu}_{bp}$	UCL (95%)
All sites	4,840	26	2.9 ± 1.1	2.0	3.3	5.1
CG/CNG sites only	670	16	13.0 ± 6.4	8.2	15.1	24.8
Non-CG/CNG sites	4,170	10	1.3 ± 0.8	0.7	1.6	2.8

[a] Mutation rate estimates for the classical approach are for $g = 8{,}000$.

[b] s.d., standard deviation; LCL, lower credible level; UCL, upper credible level.

[c] Values of $\hat{\mu}_{bp}$ for the Bayesian approach are the mean of 50,000 accepted observations and error estimates are the 95% LCL and UCL.

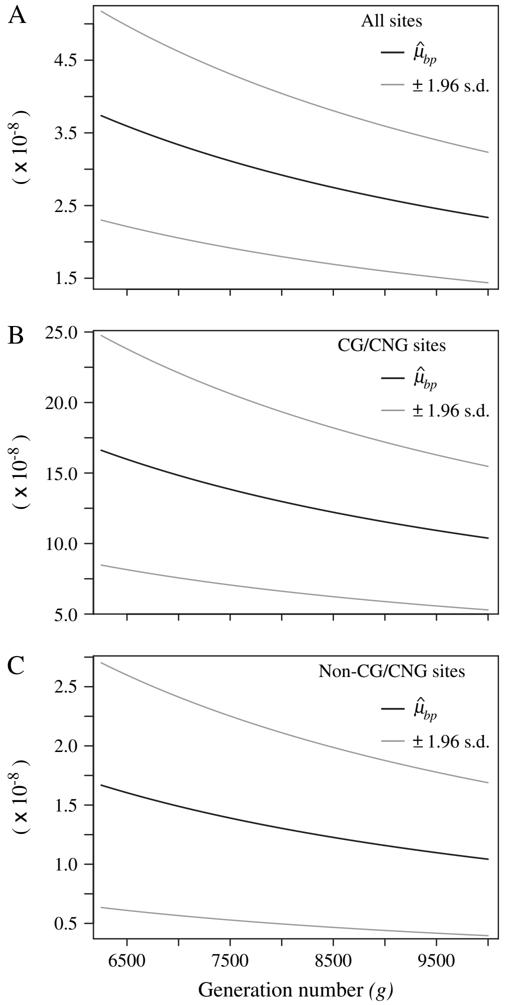

FIG. 2.—Estimates of μ_{bp} from the classical approach as a function of generation number (g) from 6,250 to 10,000. General (A), methylated (B), and nonmethylated (C) rates are shown ±1.96 standard deviations (s.d.).

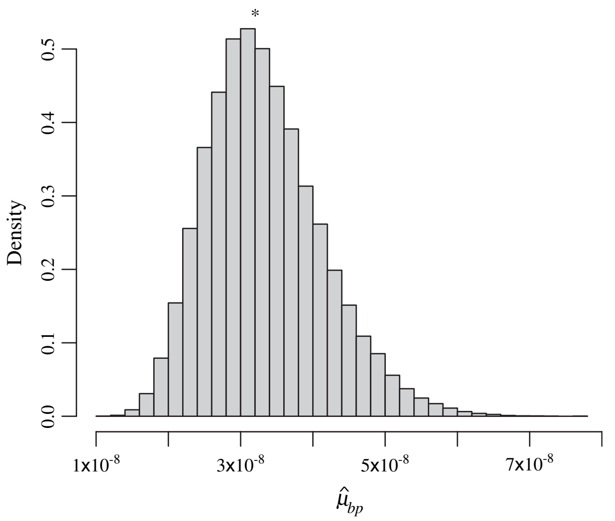

FIG. 3.—Distribution of posterior estimates for μ_{bp} using all sites (50,000 accepted observations are shown). The position of the mean (3.3×10^{-8}, see table 3) is indicated by an asterisk.

Dating of Transposon Insertion Times for Maize

A striking structural feature of the maize genome is the nested and/or tandem insertion pattern of transposable elements at intergenic locations (SanMiguel et al. 1996). Using the observation that transposable elements typically have identical terminal repeat sequences upon insertion, SanMiguel et al. (1998) used sequence divergence between terminal repeats to date the insertion time of 18 intergenic maize transposons between 0.12 and 5.15 MYA (the average insertion time was 1.76 MYA). In their dating calculations, SanMiguel et al. (1998) used the phylogenetic rate of 6.5×10^{-9} substitutions per synonymous site per year estimated from the *adh1* and *adh2* genes in grasses (Gaut et al. 1996). There are several uncertainties in applying this phylogenetic rate to maize sequence data (see *Discussion*). Also, SanMiguel et al. (1998) cautioned that the phylogenetic rate from *adh1*- and *adh2*-coding sequences may be inappropriate to apply to transposon data. In particular, they cautioned that a higher transition to transversion ratio is observed in the terminal repeat data set compared to genic sequences (SanMiguel et al. 1998). The high ratio is consistent with increased sequence evolution at intergenic transposon sequences relative to genes that may be caused by differential methylation (SanMiguel et al. 1998). In this context, it is informative that ~93% of repetitive maize sequences appear to be methylated (Palmer et al. 2003).

Because the mutational spectrum for the *tb1* intergenic region also reveals a high transition to transversion ratio (table 1), the general rate estimates for nucleotide substitution for the *tb1* intergenic region seem relevant to apply to dating maize transposition events. Applying the general Bayesian estimate from this study (3.3×10^{-8}, table 3) to the data of SanMiguel et al. (1998), the time of the most recent insertion event is 2.4×10^4, the average insertion event is 3.5×10^5, and the oldest insertion event is 1.0×10^6 years B.P. Applying the 95% lower credible value for this rate estimate (2.0×10^{-8}, table 3), the corresponding values are 4.0×10^4, 5.7×10^5, and 1.7×10^6. Applying the 95% upper credible value for this rate estimate (5.1×10^{-8}, table 3), the corresponding values are 1.6×10^4, 2.2×10^5, and 6.6×10^5.

Discussion

The low rate of nucleotide substitution makes direct estimation of μ_{bp} very difficult. In this study, we have estimated μ_{bp} for maize from intraspecific polymorphism at a major domestication locus and information about the timing and population history of maize that has come from archeological and demographic studies. While this approach addresses some sources of uncertainty associated with traditional approaches to calculating substitution rates, estimates of μ_{bp} from this study are subject to several sources of error.

A critical assumption for this study is that we correctly identified segregating sites in the *tb1* region that arose by new mutation after initial domestication. If some of the sites we use to calculate μ_{bp} did not arise by new mutation in maize but rather were introduced from teosinte populations, the estimates of μ_{bp} we report would be systematically greater than the true rates of substitution. This is a particular concern given that teosinte populations appear to have an excess of low-frequency variants relative to neutral expectations as revealed by negative values of the Tajima's D statistic (Tenaillon et al. 2004; Wright et al. 2005). In our study, we compared maize sequences in the selected region to analogous teosinte sequences where available. Although we did observe shared polymorphisms between maize and teosinte samples, analysis of teosinte haplotypes suggested that the shared polymorphisms in our data set likely resulted from new mutation in maize subsequent to domestication as opposed to introduction of polymorphism from standing genetic variation in teosinte populations. However, diversity data from teosinte for the *tb1* intergenic region is limited (sample sizes ~8, Clark et al. 2004) and may not be representative of the teosinte population or populations from which maize was initially domesticated. We cannot therefore exclude the possibility that some polymorphisms we considered as new mutations may predate domestication. We note, however, that if as few as 40%–65% of the polymorphisms we considered as new mutations in this study are used for rate calculations, the revised estimates remain within the 95% CIs for the estimates based on the complete data set (for given rate estimates using all sites, unpublished data).

One mechanism by which preexisting polymorphisms from nonselected haplotypes may have been introduced into the *tb1* upstream region is by recombination following the fixation event at *tb1*. Because a lack of recombination between the target of selection and other regions we have sampled for nucleotide polymorphism is a critical assumption, we examined the pattern of polymorphism near *tb1* to see if it is consistent with this assumption. If recombination has been frequent within the intergenic region we have examined, a general expectation is that the number of segregating sites should increase as a function of distance from a selected site somewhere in the region (as a result of recombination with divergent haplotypes). For the most part, this pattern is not observed, and the number of segregating sites is similar among regions with the exception of the 7.1- and 35.6-kb regions (where substitution frequency may have been influenced by methylation, see below). In particular, rate estimates for the 1.7- and 58.6-kb regions, which define the endpoints of the core selective sweep at *tb1*, are very low. The lack of apparent recombination is consistent with the evidence that recombination within intergenic regions in maize is uncommon (Fu, Zheng, and Dooner 2002). This finding does not exclude the possibility that an effect of recombination has biased our estimates. It does suggest, however, that such an effect (if it exists) is probably minor.

A further uncertainty in our rate estimates originates from our incomplete knowledge of the demographic history of maize. The classical approach assumes a star phylogeny, which only approximates the rapid population expansion that is thought to have occurred following initial maize domestication. We believe that the coalescent approach provides a more realistic demographic scenario by conditioning the structure of the underlying genealogy on an exponential population expansion (indeed, any other demographic scenario can in principle be incorporated into the analysis; e.g., Tavaré et al. 1997). Several previous studies have modeled maize domestication with the coalescent assuming a simple stepwise increase in population size following the initial domestication event (Eyre-Walker et al. 1998; Vigouroux et al. 2002*b*). One advantage of the present coalescent approach is that it should allow revision of $\hat{\mu}_{bp}$ as additional information about the population history of maize becomes known.

Effect of Methylation on Mutation Pattern and Rate

In maize, transposable elements and other repetitive sequences are hypermethylated relative to genic DNA (Palmer et al. 2003). In this context, the high frequency of mutations at sites of symmetric C methylation (CG and CNG) in the *tb1* intergenic region is not surprising given that the *tb1* upstream region has a large number of repetitive sequences (Clark et al. 2004). However, the structure of intergenic regions in grasses is complex, and large tracts of repetitive DNA are often separated by nonrepetitive regions that appear to be nongenic (e.g., Song, Llaca, and Messing 2002). In fact, the six regions examined in this study are largely unique (Clark et al. 2004), and relatively little is known about the methylation state at such unique, nongenic sequences interspersed among tracts of repetitive maize DNA. It is striking that the strong bias for mutation at CG and CNG sites is primarily observed at the 7.1- and 35.6-kb regions. This may indicate an uneven pattern of methylation between sequences located 5′ to the *tb1*-coding region.

Because maize genes are largely (but not exclusively) devoid of methylation (Palmer et al. 2003), and because most existing polymorphism data for maize come from genic sequences, we calculated a "nonmethylated" mutation rate from our data with the expectation that this rate may reflect the rate of mutation in genic sequences. In calculating nonmethylated rates, we simply excluded mutations at CG and CNG sites from analysis. In general, the nonmethylated rates for the *tb1* region were about an order of magnitude lower than the rates we calculated for CG and CNG sites alone. A similar effect has been noted in humans, in which the rate of mutation at CG sites (the target of symmetric methylation in mammals) is an order of magnitude higher than the rate at other sites (Nachman and Crowell 2000). A caveat for our nonmethylated rate estimates is that asymmetric methylation at C residues does occur in plants (Finnegan et al. 1998), and we have no straightforward way to relate asymmetric methylation to mutational frequency. It is striking, however, that at the 7.1- and 35.6-kb regions, 15 of 16 transitions occurred at consensus sites of symmetric methylation, even though C residues located within CG and CNG sites represented only 29% of all C residues at these regions. The paucity of mutations at non-CG/CNG sites suggests that asymmetric methylation has not had a comparatively large effect on mutation rates in the region. Also, the 1:1 transition to transversion ratio that is observed if transitions at CG and CNG sites are excluded from the analysis is more typical of that found in genic regions in maize (e.g., *adh1* intron sequences, SanMiguel et al. 1998). This lends some credence to applying nonmethylated rates from this study to existing genic data sets.

Comparison to Other Rate Estimates

It is informative to compare the rates from our study to previous estimates of μ_{bp} for grasses. Gaut et al. (1996) calculated rates of 7.0×10^{-9} and 6.0×10^{-9} substitutions per synonymous site per year for the grass *adh1* and *adh2* genes, respectively, assuming a divergence time of 50 Myr between maize and rice/barley. Extending this work, White and Doebley (1999) used a maize/rice comparison to calculate synonymous rates for five additional genes (*ant*, *c1*, *c2*, *cdc*, and *ohp*; Gaut and Doebley, 1997) and obtained an average rate of 5.9×10^{-9} when the data from *adh1* and *adh2* were included. Applying these rates to maize data has several shortcomings (see White and Doebley 1999 for a detailed discussion). One concern is the uncertainty in the fossil estimate of the maize-rice divergence time (40–70 MYA in different studies; see Wolfe, Sharp, and Li 1989; Gaut et al. 1996). In addition, Gaut and Clegg (1993) have reported that the rate of sequence evolution at *adh1* is 1.7 times as high in the maize lineage as in the pennisetum lineage. If an accelerated rate of sequence evolution is found in the maize lineage, then general grass rates may be inappropriate to apply to maize (White and Doebley 1999).

In all cases, estimates of μ_{bp} from this study are higher than the previously reported grass estimates. The nonmethylated rate estimates from this study are probably most relevant to compare to phylogenetic estimates based on genic sequences (see above) and are ~two- to threefold higher. A similar discrepancy has been reported for rate estimates for mitochondrial evolution in Adélie penguins (*Pygoscelis adeliae*), in which estimates of substitution rates that incorporate information from ancient DNA sequences are two to seven times higher than estimates that come from indirect phylogenetic rate approaches (Lambert et al. 2002).

The absolute values for rate estimates from this study are similar to those reported from several studies that have examined sequence evolution in lineages with comparatively recent and well characterized divergence times. For example, Nachman and Crowell (2000) estimated a rate of 2.5×10^{-8} substitutions per site per year in a human-chimpanzee comparison. In plants, Koch, Haubold, and Mitchell-Olds (2000) used an estimate of divergence time between *Barbarea* and *Cardamine* based on fossil pollen data to derive a substitution rate of 1.5×10^{-8} for the *chs* and *adh* genes.

Transposon Expansion and Genome Evolution in Maize

The large size of the maize genome relative to rice or sorghum can be explained partly by the expansion of retrotransposon families in the *Zea* lineage (SanMiguel et al. 1996). In a previous study, SanMiguel et al. (1998) used divergence between terminal repeat sequences and a phylogenetic estimate of μ_{bp} to infer that the increase in transposon copy number in the maize genome occurred largely within the last 5 Myr. However, SanMiguel et al. (1998) noted the tentative nature of this conclusion and suggested that the expansion may have been more recent if the phylogenetic rate used in their study is an underestimate of the true mutation rate for transposon sequences. We have repeated the calculations of SanMiguel et al. (1998) using a general rate estimate for the *tb1* intergenic region that we believe is the most appropriate current estimate of μ_{bp} to apply to dating transposon insertion events in maize.

Our calculations confirm the suspicions of SanMiguel et al. (1998) that expansion of transposon families in the *Z. mays* genome has occurred very recently on an evolutionary timescale. This finding is consistent with a growing body of evidence that suggests that expansion of transposable element populations has occurred within the last ~10 Myr in both monocot and dicot species (see Bennetzen, Ma, and Devos 2005, and references therein). The most recent of the 18 maize transposition events we have dated is predicted to have occurred only 24,000 years ago and thus nearly overlaps the beginning of maize domestication from ssp. *parviglumis* by humans. This finding is consistent with several experimental observations in maize. First, it is known that transposition is ongoing in some maize germplasm (e.g., Varagona, Purugganan, and Wessler 1992). Second, recent genome sequencing projects have revealed substantial polymorphism for the presence or absence of sequences between maize alleles, including transposable elements (Fu and Dooner 2002). This finding is consistent with recent transposition in the maize genome. As the causal polymorphisms that underlie phenotypic alterations selected during domestication are discovered, it will be exciting to determine if the movement of transposable elements has contributed to the pool of genetic variation tapped by early maize agriculturalists.

Acknowledgments

We would like to thank Y. Vigouroux and A.-C. Thuillet for helpful discussions. We are especially appreciative of two anonymous reviewers and the associate editor for comments that have led to improvements to the manuscript. This work was supported by National Institutes of Health Awards F32 GM-65008 (to R.M.C.) and GM058816 (to J.D.). S.T. is a Royal Society/Wolfson Research Merit Award holder.

Literature Cited

Bennetzen, J. L., J. Ma, and K. M. Devos. 2005. Mechanisms of recent genome size variation in flowering plants. Ann. Bot. **95**: 127–132.

Brinkmann, B., M. Klintschar, F. Neuhuber, J. Huhne, and B. Rolf. 1998. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am. J. Hum. Genet. **62**:1408–1415.

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.

Coulondre, C., J. H. Miller, P. J. Farabaugh, and W. Gilbert. 1978. Molecular basis of base substitution hotspots in *Escherichia coli*. Nature **274**:775–780.

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

———. 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 C. Gustus. 1995. *teosinte branched1* and the origin of maize: evidence for epistasis and the evolution of dominance. Genetics **141**:333–346.

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

Doganlar, S., A. Frary, M. C. Daunay, R. N. Lester, and S. D. Tanksley. 2002. Conservation of gene function in the solanaceae as revealed by comparative mapping of domestication traits in eggplant. Genetics **161**:1713–1726.

Drake, J. W., B. Charlesworth, D. Charlesworth, and J. F. Crow. 1998. Rates of spontaneous mutation. Genetics **148**:1667–1686.

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.

Finnegan, E. J., R. K. Genger, W. J. Peacock, and E. S. Dennis. 1998. DNA methylation in plants. Annu. Rev. Plant Physiol. Plant Mol. Biol. **49**:223–247.

Fu, H., and H. K. Dooner. 2002. Intraspecific violation of genetic colinearity and its implications in maize. Proc. Natl. Acad. Sci. USA **99**:9573–9578.

Fu, H., Z. Zheng, and H. K. Dooner. 2002. Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. Proc. Natl. Acad. Sci. USA **99**:1082–1087.

Gaut, B. S., and M. T. Clegg. 1993. Nucleotide polymorphism in the *Adh1* locus of pearl millet (*Pennisetum glaucum*) (Poaceae). Genetics **135**:1091–1097.

Gaut, B. S., and J. Doebley. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. Proc. Natl. Acad. Sci. USA **94**:6809–6814.

Gaut, B. S., B. R. Morton, B. C. McCaig, and M. T. Clegg. 1996. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. Proc. Natl. Acad. Sci. USA **93**:10274–10279.

Griffiths, R. C., and S. Tavaré. 1994. Sampling theory for neutral alleles in a varying environment. Phil. Trans. R. Soc. Lond. B **344**:403–410.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

Koch, M. A., B. Haubold, and T. Mitchell-Olds. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). Mol. Biol. Evol. **17**:1483–1498.

Kumar, S., and S. Subramanian. 2002. Mutation rates in mammalian genomes. Proc. Natl. Acad. Sci. USA **99**:803–808.

Lambert, D. M., P. A. Ritchie, C. D. Millar, B. Holland, A. J. Drummond, and C. Baroni. 2002. Rates of evolution in ancient DNA from Adelie penguins. Science **295**:2270–2273.

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

Nachman, M. W., and S. L. Crowell. 2000. Estimate of the mutation rate per nucleotide in humans. Genetics **156**:297–304.

Palmer, L. E., P. D. Rabinowicz, A. L. O'Shaughnessy, V. S. Balija, L. U. Nascimento, S. Dike, M. de la Bastide, R. A. Martienssen, and W. R. McCombie. 2003. Maize genome sequencing by methylation filtration. Science **302**:2115–2117.

Piperno, D. R., and K. V. Flannery. 2001. The earliest archaeological maize (Zea mays L.) from highland Mexico: new accelerator mass spectrometry dates and their implications. Proc. Natl. Acad. Sci. USA **98**:2101–2103.

SanMiguel, P., B. S. Gaut, A. Tikhonov, Y. Nakajima, and J. L. Bennetzen. 1998. The paleontology of intergene retrotransposons of maize. Nat. Genet. **20**:43–45.

SanMiguel, P., A. Tikhonov, Y. K. Jin et al. (11 co-authors). 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science **274**:765–768.

Slatkin, M., and R. R. Hudson. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129**:555–562.

Smith, B. D. 1998. The emergence of agriculture. W H Freeman & Co., New York.

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

Song, R., V. Llaca, and J. Messing. 2002. Mosaic organization of orthologous sequences in grass genomes. Genome Res. **12**:1549–1555.

Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly. 1997. Inferring coalescence times for molecular sequence data. Genetics **145**:505–518.

Tenaillon, M. I., M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley, and B. S. Gaut. 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.

Varagona, M. J., M. Purugganan, and S. R. Wessler. 1992. Alternative splicing induced by insertion of retrotransposons into the maize waxy gene. Plant Cell **4**:811–820.

Vigouroux, Y., J. S. Jaqueth, Y. Matsuoka, O. S. Smith, W. D. Beavis, J. S. Smith, and J. Doebley. 2002*a*. 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, S. Kresovich, Y. Matsuoka, and J. Doebley. 2002*b*. 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, R. L., A. Stec, J. Hey, L. Lukens, and J. Doebley. 1999. The limits of selection during maize domestication. Nature **398**:236–239.

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

Wolfe, K. H., P. M. Sharp, and W.-H. Li. 1989. Rates of synonymous substitution in plant genes. J. Mol. Evol. **29**:208–211.

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

Brandon Gaut, Associate Editor

Accepted July 26, 2005