Evolution of Anthocyanin Biosynthesis in Maize Kernels: The Role of Regulatory and Enzymatic Loci

Michael A. Hanson,*[,1] Brandon S. Gaut,[†] Adrian O. Stec,* Susan I. Fuerstenberg,* Major M. Goodman,[‡] Edward H. Coe[§] and John F. Doebley*

**Department of Plant Biology, University of Minnesota, St. Paul, Minnesota 55108, [†]Center for Theoretical and Applied Genetics, Department of Plant Sciences, Rutgers University, New Brunswick, New Jersey 08903, [‡]Department of Crop Science, North Carolina State University, Raleigh, North Carolina 27695 and [§]USDA-ARS and Department of Agronomy, University of Missouri, Columbia, Missouri 65211*

Manuscript received October 28, 1995
Accepted for publication April 10, 1996

ABSTRACT

Understanding which genes contribute to evolutionary change and the nature of the alterations in them are fundamental challenges in evolution. We analyzed regulatory and enzymatic genes in the maize anthocyanin pathway as related to the evolution of anthocyanin-pigmented kernels in maize from colorless kernels of its progenitor, teosinte. Genetic tests indicate that teosinte possesses functional alleles at all enzymatic loci. At two regulatory loci, most teosintes possess alleles that encode functional proteins, but ones that are not expressed during kernel development and not capable of activating anthocyanin biosynthesis there. We investigated nucleotide polymorphism at one of the regulatory loci, *c1*. Several observations suggest that *c1* has not evolved in a strictly neutral manner, including an exceptionally low level of polymorphism and a biased representation of haplotypes in maize. Curiously, sequence data show that most of our teosinte samples possess a promoter element necessary for the activation of the anthocyanin pathway during kernel development, although genetic tests indicate that teosinte *c1* alleles are not active during kernel development. Our analyses suggest that the evolution of the purple kernels resulted from changes in *cis* regulatory elements at regulatory loci and not changes in either regulatory protein function nor the enzymatic loci.

DISCERNING the type of genes involved in adaptive evolution and the nature of the alterations in these genes are fundamental challenges for evolutionary biology. Several authors have proposed that the evolution of new phenotypes more often involves changes in regulatory genes than in the numerous downstream genes under their control (GOODRICH *et al.* 1992; DOEBLEY 1993). This is an especially attractive model for macroevolutionary changes because it allows for the coordinate activation of multiple downstream genes via a single (or small number) change in their upstream regulator(s), which seems more probable than multiple independent changes at each downstream gene. Testing this model requires study systems in which both regulatory and target genes are known and in which one can use genetic approaches to identify the genes involved in phenotypic evolution.

To investigate the roles of regulatory *vs.* target genes in evolutionary change, we have chosen the anthocyanin biosynthetic pathway of maize (*Zea mays* L. ssp. *mays*). Several features make this an attractive system for evolutionary study. First, this pathway controls a phenotype, anthocyanin-pigmented kernels, that is common among maize landraces but unknown in the wild progenitor of maize, teosinte (*Zea* spp.). The evolution of anthocyanin-pigmented kernels is most certainly the result of human selection that favored the brilliant hues of purple, red, and blue that anthocyanin produces over the far less colorful anthocyaninless kernels of teosinte. Second, the structural and regulatory loci in the pathway have been characterized at both the molecular and genetic levels. Thus, one can simultaneously assay changes at the different levels of the regulatory hierarchy. Third, because maize and teosinte are fully interfertile, one can use genetics to identify the genes that control the phenotypic differences between them.

The anthocyanin pathway of maize includes eight known enzymatic genes (*a1, a2, bz1, bz2, c2, chi, pr* and *whp*) that catalyze the biosynthesis or transport of anthocyanin, and five regulatory genes (*b, c1, pl, r,* and *vp1*) that govern the tissue-specific expression of anthocyanin synthesis (COE *et al.* 1988; HOLTON and CORNISH 1995). Available genetic stocks enable one to determine the allelic composition of a new type of maize or teosinte at most of the above loci.

Among the loci in the anthocyanin pathway, we have chosen *c1* for molecular evolutionary analysis. The

Corresponding author: John Doebley, Department of Plant Biology, University of Minnesota, St. Paul, MN 55108.
E-mail: doebley@maroon.tc.umn.edu

[1] *Present address:* Division of Mathematics and Natural Sciences, Elmira College, Elmira, NY 14902.

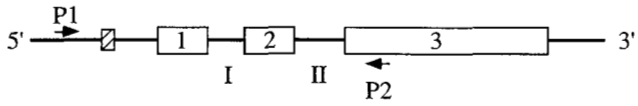

FIGURE 1.—A schematic representation of the *c1* gene. Exons are numbered and shown as wide unshaded boxes. Introns are represented by roman numerals. The hatched box denotes the promoter region containing Boxes I and II. P1 and P2 refer to the primers used in PCR amplifications (see MATERIALS AND METHODS).

structure and function of the *c1* locus have been studied in remarkable detail, and a number of *c1* alleles have been characterized at both genetic and molecular levels. The wild-type *C1* allele contains three exons and two introns (Figure 1) and encodes a protein of 273 amino acids in length (PAZ-ARES *et al.* 1987). The wild-type protein functions as a transcriptional activator of structural genes in the anthocyanin pathway (PAZ-ARES *et al.* 1990; GOFF *et al.* 1991). Critical to our study, functional assays have identified the elements in the *c1* promoter necessary for its expression during kernel development (HATTORI *et al.* 1992).

Study of evolutionary questions in maize and teosinte is facilitated by extensive research on their systematics (WILKES 1967; KATO 1976; GOODMAN 1978; ILTIS and DOEBLEY 1980; MCCLINTOCK *et al.* 1981; ZIMMER *et al.* 1988). The genus Zea consists of four species of outcrossing perennial and annual grasses native to Mexico and Central America. The genus is divided into two sections: section *Luxuriantes,* which contains *Z. diploperennis, Z. perennis,* and *Z. luxurians,* and section *Zea,* which contains a single highly polymorphic species, *Z. mays.* DOEBLEY (1990a) defined four subspecies within *Z. mays:* (1) ssp. *parviglumis,* a teosinte from southwestern Mexico, (2) ssp. *mexicana,* a teosinte from central and northern Mexico, (3) ssp. *huehuetenangensis,* a teosinte from western Guatemala, and (4) ssp. *mays,* the cultigen maize, which probably shares its closest relationship to ssp. *parviglumis* (APPENDIX).

In this article, we take a two pronged approach to understanding how the anthocyanin-pigmented kernels of maize evolved and how selection for this trait may have acted on *c1*. First, we use genetic tester stocks to ask whether teosinte lacks the necessary regulatory or enzymatic functions needed to produce purple kernels. Second, we examine nucleotide polymorphism at *c1* in maize and teosinte to ask if maize and teosinte differ for known *cis*-regulatory elements and how selection and other forces have affected the pattern of polymorphism at *c1*. Results from these two approaches suggest that the evolution of anthocyanin-pigmented kernels in maize resulted largely from changes at the regulatory loci including selective elimination of one class of *c1* haplotypes from the maize gene pool.

MATERIALS AND METHODS

Allelic diversity: Allelic diversity at the loci in the anthocyanin pathway was determined by crossing a diverse selection of teosintes onto tester stocks for *a1, a2, bz1, bz2, c1, c2, pr* and *r*. Each tester stock carried a recessive allele at one of these loci and dominant alleles at the other loci in the pathway. All testers also carried the recessive alleles at *b1, pl1* and *y1,* and all were constructed in the hybrid W23 × K55. For each test, a teosinte sample plant was used as the pollen parent and the tester as the female parent. In all tests, if the teosinte sample plant carried a functional (dominant) allele at the test locus, then purple kernels should develop on the tester stock. Alternatively, if the teosinte carried a recessive allele at the test locus, then white kernels should develop, except for the *pr* tester, for which *pr* was indicated by red kernels.

The *c1* locus has several genetically defined allelic variants: *C1,* the functional or wild-type allele; *C1-I,* a dominant inhibitor of transcriptional activation; *c1-n,* a recessive null allele; and *c1-p,* a recessive allele expressed only during seed germination and only in the presence of light (COE *et al.* 1988). To distinguish between white kernels resulting from *C1-I* and white kernels resulting from the recessive alleles, white kernels from the test crosses were grown and the plants selfed and backcrossed to a tester with purple kernels. If *C1-I* was present in the teosinte sample plant, then 50% of the kernels on the purple kernel tester will be white, whereas if a recessive allele was present, then 0% of the kernels will be white. Similarly, kernels on the selfed plants will be 3/16 purple if *C1-I* is present *vs.* 9/16 purple if there was a recessive. These figures are 3/16 and 9/16 rather than 1/4 and 3/4 because all teosintes also possess *r* rather than *R.* Recessive *c1* alleles were further classified as *c1-p* (positive) or *c1-n* (negative) by germinating the kernels from crosses with the *c1* tester stock in the presence of light (CHEN and COE 1977). *c1-p* induces the production of anthocyanin during germination with light, while *c1-n* does not.

The *r* locus has both kernel and plant components with the following designations: *r-g* (white kernels-green plant), *R-g* (purple kernels-green plant), *r-r* (white kernels-red plant) and *R-r* (purple kernels-red plant) (COE *et al.* 1988). The test described above assayed the kernel component of *r*. To score the plant component, kernels from the test crosses on the *r* tester (*r-g*) were grown to small seedlings. Red color on the coleoptile or tip of the first leaf indicates the presence of a functional plant component, while purely green seedlings indicate its absence.

Sequence sampling: Sequences of the *c1* locus from 26 individuals were isolated by PCR (Table 1). Because teosinte and maize are highly polymorphic, outcrossing plants, it was anticipated that many individual plants from natural populations of teosinte and open-pollinated maize landraces would be heterozygous at *c1*. DNA from heterozygous plants is less suitable than DNA from homozygous plants as a substrate for PCR since it contains a mixture of two haplotypes that will confuse reading autoradiographs and may result in inaccurate sequence data. For this reason, two strategies were employed to obtain DNA samples possessing only a single *c1* haplotype. The "F_2" method involved crossing plants of the maize and teosinte sample populations to a maize tester stock. A single F_1 plant from each cross was self-pollinated to generate an F_2 population. Southern hybridization analysis of F_2 individuals using a cloned portion of *c1* and low copy number sequences (*umc105, umc113* and *bz1*) that flank *c1* enabled the identification of plants homozygous for the *c1* sample allele. The "gel" isolation method involved digestion of DNAs from sample plants with *Bam*HI, which does not cut in *c1,* and Southern hybridization using a clone of *c1* as the probe. This procedure enabled us to identify heterozygous individuals and to excise restriction fragments from an electrophoretic gel that possessed only a single *c1* haplotype. In addition to these two

TABLE 1

Taxa and collections analyzed for *c1* nucleotide sequence

Taxon	Race	Country	Collection[a]	Method[b]	Haplotype
Z. mays ssp. *mays*	U.S. Inbred	USA	W22-LC[c]	—	1
Z. mays ssp. *mays*	Hickory King	USA	PI: 311237	Gel	1
Z. mays ssp. *mays*	U.S. hybrid	USA	W23 × K55	F_2	2
Z. mays ssp. *mays*	Jala	Mexico	G: Jal 42	Gel	3
Z. mays ssp. *mays*	Olotillo	Mexico	G: Chs 56	Gel	3
Z. mays ssp. *mays*	Harinoso de Ocho	Mexico	G: Nay 24	Gel	4
Z. mays ssp. *mays*	Pira	Venezuela	G: Ven 485	Gel	5
Z. mays ssp. *mays*	—	Peru	H: 1-468[d]	Total	6
Z. mays ssp. *mays*	Assiniboine	USA	PI: 213793	Gel	6
Z. mays ssp. *mays*	Enano Gigante	Ecuador	G: Ecu 969	Gel	7
Z. mays ssp. *mays*	Corioco	Bolivia	G: Bov 396	Gel	7
Z. mays ssp. *mays*	Altiplano	Bolivia	G: Bov 903	Gel	7
Z. mays ssp. *mays*	Acoma Pueblo	USA	PI: 218167	Gel	8
Z. mays ssp. *mexicana*	Chalco	Mexico	D: 479	F_2	1
Z. mays ssp. *mexicana*	Central Plateau	Mexico	K: 67-22	F_2	7
Z. mays ssp. *mexicana*	Chalco	Mexico	D: 482	F_2	8
Z. mays ssp. *mexicana*	Chalco	Mexico	I&D: 401	Gel	10
Z. mays ssp. *mexicana*	Nobogame	Mexico	B: Nobogame	Gel	15
Z. mays ssp. *parviglumis*	Balsas	Mexico	K: 77-13	Gel	1
Z. mays ssp. *parviglumis*	Balsas	Mexico	P: 11065	F_2	11
Z. mays ssp. *parviglumis*	Balsas	Mexico	I&C: 81	Gel	12
Z. mays ssp. *parviglumis*	Balsas	Mexico	K: 67-20	Gel	13
Z. mays ssp. *parviglumis*	Balsas	Mexico	C: 10-78	F_2	16
Z. luxurians	Guatemala	Guatemala	I: G-36	Total	7
Z. luxurians	Guatemala	Guatemala	I: 30900	Total	9
Z. diploperennis	—	Mexico	I: 1190	F_2	1
Z. diploperennis	—	Mexico	I: 2549	Total	14

[a] Collections are designated by a single letter for the collector/curator followed by a collection designation. Collectors/curators are as follows: B, Beadle; C, Cervantes; D, Doebley; G, Goodman; H, Hastorf; I, Iltis; I&C, Iltis and Cochrane; I&D, Iltis and Doebley; K, Kato; P, Puga; PI, USDA Plant Introduction Station.

[b] The method by which DNA for PCR was isolated (see MATERIALS AND METHODS).

[c] Sequence from PAZ-ARES *et al.* (1987).

[d] DNA for this sequence was extracted from a partially carbonized archaeological maize kernel from the Pancan site, Junin Province, Peru. This kernel (ID No. 1986-2555) was provided to us by Dr. CHRISTINE HASTORF and dates between 550 and 900 A.D.

strategies, three sequences were isolated by PCR amplification with "total" DNA of individual teosinte plants grown from seed collected from natural populations, and one sequence was amplified from DNA extract from an archaeological kernel (Table 1). In sampling the taxa, an attempt was made to represent their geographic diversity.

PCR primers were designed to amplify much of the promoter region of the *c1* locus and some of the coding region (Figure 1). The 5′ primer (CACTGGGGATCCTTAGTTACTGGCATG) was designed to hybridize to a position 380 bp upstream of the ATG start codon. The 3′ primer (CATAGGTACCAGCGTGCTGTTCCAGTAGT) was made specific to sequence in the third exon, ~580 bp downstream from the ATG start codon. These primers contain *Bam*HI (5′ primer) and *Kpn*I (3′ primer) restriction sites that were used for cloning. The PCR-amplified molecules were ligated into the vector pUC19 and transformed into the *Escherichia coli* strain DH-5α. For sample sequences isolated by PCR of DNAs from the gel and F_2 methods, 10 or more independent pUC clones were obtained and pooled for DNA sequencing (Sequenase version 2.0, U.S. Biochemical, Inc.). The pooling strategy was employed to reduce the chance of introducing PCR-generated errors into the sequence. Any single clone might contain a PCR error; however, the same error is unlikely to occur in any two independent clones. Thus, in the pool of 10 or more independent clones, any PCR errors will be represented in 10% or less of the template for the sequencing reaction and should not be visible on the autoradiograph. All sequences were determined in both directions using a series of internal primers.

A number of maize *c1* sequences are available in the Genbank data base. Most of these sequences were isolated because of the mutant phenotype they confer, and so most were not included in our "random" sample of alleles. However, we did include one previously published sequence in our sample, the dominant wild-type allele, *C1* (PAZ-ARES *et al.* 1987). With this sequence, our maize sample of *c1* consists of 13 sequences (Table 1).

Sequence analysis: Phylogenetic reconstruction was performed by the neighbor-joining method (SAITOU and NEI 1987), using the KIMURA (1980) two-parameter model to estimate distances between sequences. The data were resampled 500 times for bootstrap analyses. Nucleotide diversity was summarized by the statistic $\hat{\theta}$ (WATTERSON 1975) and confidence intervals around $\hat{\theta}$ were calculated using the recursion method of KREITMAN and HUDSON (1991).

Tests for selection were based on the methods of TAJIMA (1989) and HUDSON *et al.* (1987) (the HKA test). The latter were performed using average pairwise differences as the measure of divergence. The probability of observing the HKA statistic under the null hypothesis of neutrality was determined by simulation of the coalescent process, using the program of HILTON *et al.* (1995). Simulations were a parametric bootstrap into which estimates of the parameters f, θ, and T were incorporated (see HUDSON *et al.* 1987), and probabilities were based on 2000 parametric bootstraps. Only third position substitutions and substitutions within intron and flanking regions were considered in HKA tests.

We used the pairwise method of GAUT and CLEGG (1993) to test for homogeneity of substitution rates along nucleotide sequences. This method uses a likelihood ratio statistic to test for homogeneity in s, the number of nucleotide substitutions per nucleotide site, along the length of a gene. This test requires the *a priori* partitioning of sequences into regions. In this study, genic regions were defined by intron/exon boundaries.

Tests for genetic subdivision between samples were based on the resampling procedure described by HUDSON *et al.* (1992a). We employed the K_s and K_s* statistics of HUDSON *et al.* (1992a), using pairwise KIMURA (1980) two-parameter distances as the measure of differences between sequences. The statistics K_s and K_s* gave very similar results; we report the results of the subdivision tests that used the K_s* statistic. Test results are based on 1000 random permutations. A significant result ($P < 0.05$) indicates that taxa are genetically subdivided such that sequences within taxa are significantly more closely related to one another than are sequences between taxa.

RESULTS

Allelic diversity in *c1* and other loci: We surveyed allelic diversity in teosinte at six enzymatic (*a1, a2, bz1, bz2, c2,* and *pr*) and two regulatory (*c1* and *r*) loci in the anthocyanin pathway (Table 2). Dominant or functional alleles predominate at all enzymatic loci, although recessive alleles were observed at *a2, bz1* and *pr* in low frequency, 0.07, 0.02 and 0.04, respectively. This was the expected result since teosinte produces anthocyanin in various vegetative tissues (J. DOEBLEY, personal observation). The results for the regulatory loci are strikingly different in that the recessive or nonfunctional alleles predominate (Table 2). At *r*, we observed only the recessive kernel component among the 45 plants tested. At *c1*, the recessive allele was most common (0.84); however, we also observed the dominant *C1* (0.05) and the dominant inhibitor *C1-I* (0.11). *C1-I* had a restricted distribution occurring at a moderately high frequency in ssp. *mexicana* (0.23), while it was absent from all other taxa except one collection of ssp. *parviglumis*. Moreover, within ssp. *mexicana, C1-I* was found only in collections from the Valley of Mexico (Race Chalco teosinte) and it was in all collections from this region.

The recessive *c1* can be subclassified into two forms: *c1-p* (positive) for kernels that turn purple when germinated in the presence of light and *c1-n* (negative) for kernels that remain white when germinated in the presence of light (CHEN and COE 1977). We tested 35 of the *c1* alleles and found both *c1-p* and *c1-n* at frequencies of 0.51 and 0.49, respectively. Both *c1-p* and *c1-n* were present in ssp. *parviglumis* and ssp *mexicana*. The *r* locus has a second component that regulates anthocyanin production in vegetative tissues (COE *et al.* 1988). We tested 40 *r* alleles and found both *r-r* (white seed; red plant) and *r-g* (white seed; green plant) at frequencies of 0.97 and 0.03, respectively.

Nucleotide polymorphism in the *c1* locus: The sample of a single sequence from each of 27 individuals resulted in 16 distinct haplotypes (Table 1; Figure 2). Haplotypes 1, 7 and 8 were found in more than one individual, with haplotypes 1 and 7 found most frequently. Haplotype 1, the wild-type allele, was found in maize, ssp. *parviglumis,* ssp. *mexicana,* and *Z. diploperennis*. Haplotype 7 was found in maize and *Z. luxurians,* and haplotype 8 was found in maize and ssp. *mexicana.*

Figure 2 provides a summary of the polymorphism found in *c1*. Two nucleotide substitutions result in amino acid replacements, and both of these replacements are conservative with regard to charge and hydrophobicity. For example, the amino acid replacement in haplotype 11 substitutes neutral and hydrophobic tryptophan with neutral and hydrophobic leucine, and the amino acid replacement in haplotypes 3 and 4 substitutes an acidic residue (glutamic acid) with another acidic residue (aspartic acid). Three synonymous substitutions are found in the third position of codons; the remainder of polymorphic sites are in the introns and 5′ region.

Insertion and deletion (indel) polymorphisms are found throughout noncoding regions of *c1,* but the insertion and deletion polymorphisms in the promoter region are particularly interesting. SCHEFFLER *et al.* (1994) noted indels in two regions of the promoter and denoted them Box I and Box II (Figure 2). In a study of functional differences among *c1* haplotypes, HATTORI *et al.* (1992) demonstrated that the deletion of the "gtgtc" motif in Box I inhibits *c1* expression during seed maturation. In our study, this Box I deletion was found only in haplotypes from wild taxa. Indels in Box II also affect expression. In our survey, a Box II motif like that found in the *c1-n* and *c1-p* alleles was found in haplotype 12, and a new Box II motif was found in haplotype 10. In addition, a new insertion was found between Boxes I and II in haplotype 7.

Class I and Class II haplotypes: Examination of Figure 2 shows that the haplotypes found in maize (haplotypes 1–8), together with haplotypes 9 and 10, have very few differences between them. There are at most seven differences between haplotypes within this group. On the other hand, comparison of this group to the remaining haplotypes (12–16) reveals a great many more

TABLE 2

Anthocyanin loci in teosinte

Collection[a]	*A1*	*a1*	*A2*	*a2*	*Bz1*	*bz1*	*Bz2*	*bz2*	*C1*	*C1-I*	*c1*	*C2*	*c2*	*Pr*	*pr*	*R-g R-r*	*r-g r-r*
Zea diploperennis																	
I: 1190	1	0	1	0	1	0	2	0	0	0	2[b]	1[c]	0	2	0	0	3
Zea luxurians																	
B: Progresso	1	0	1	0	1	0	1	0	0	0	1	1[c]	0	1	0	0	1
I: G-5	2	0	2	0	2	0	2	0	0	0	2	2[c]	0	2	0	0	1
I: G-36	3	0	3	0	4	0	1	0	0	0	2	3[c]	0	3	0	0	3
I: G-38	1	0	1	0	1	0	1	0	0	0	2	1[c]	0	1	0	0	1
I: G-42	2	0	2	0	1	0	2	0	0	0	3	2[c]	0	2	0	0	3
Zea mays ssp. *huehuetenangensis*																	
I: G-120	1	0	1	0	1	0	1	0	0	0	1	1	0	3	0	0	2
Zea mays ssp. *parviglumis*																	
B&K: 1	3	0	1	0	2	0	1	0	0	0	3	2[c]	0	3	0	0	3
B&K: 4	1	0	1	0	1	0	2	0	0	0	1	1	0	2	0	0	1
B&K: 6	1	0	1	0	1	0	1	0	0	0	2	1[c]	0	1	0	0	1
C: 10-78	2	0	3	0	2	0	1	0	0	1	1	2[c]	0	3	0	0	2
C-O: Ejutla	2	0	2	0	2	0	2	0	0	0	3	2[c]	0	3	0	0	4
K: 67-15	1	0	2	0	1	0	1	0	0	0	1	1[c]	0	1	0	0	1
K: 67-20	—	—	1	0	1	0	2	0	0	0	2	1[c]	0	1	0	0	3
P: 11065	1	0	1	0	1	0	1	0	0	0	1	2[c]	0	1	1	0	2
PI: 384063	4	0	1.5	0.5	1.5	0.5	2	0	0	0	2	1[c]	0	4	0	0	2
B: Salado	1	0	2	0	1	0	1	0	0	—	1	1[c]	0	—	—	0	2
Zea mays ssp. *mexicana*																	
C: 18-78	—	—	—	—	—	—	1	0	2.5	1	0.5	1	0	—	—	—	—
D: 479	2	0	2	1	1	0	3	0	0	1	1	1	0	2	0	0	2
D: 481	3	0	2	0	1	0	1	0	0	0.5	1.5	2[c]	0	2	0	0	1
D: 482	1	0	1	0	—	—	1	0	0	0.5	1.5	—	—	1	0	0	2
D: 625	2	0	—	—	2	0	1	0	0	0	1	—	—	1	0	—	—
D: 642	3	0	1	0	—	—	1	0	—	1	—	3	0	1	0	—	—
D: 643	—	—	—	—	—	—	1	0	0	—	1	—	—	—	—	—	—
I&D: 401	1	0	—	—	—	—	—	—	0	0.5	1.5	1	0	—	—	—	—
K: 67-22	1	0	1	0	1	0	1	0	0	0	1	1	0	1	0	0	1
B: Nobogame	2	0	0	1	2	0	1	0	0	0	3	2[c]	0	1	1	0	3
P: 11066	1	0	—	—	1	0	1	0	0	0	1	2	0	1	0	0	1

Numbers in the table refer to the number of teosinte plants with a particular allele. 0.5 was used in cases where a teosinte (pollen) plant was heterozygous, resulting in ears from the test cross with half colored and half colorless kernels.

[a] Collections are designated by a single letter for the collector/curator followed by a collection designation. Collectors/curators are as follows: B, Beadle; C, Cervantes; C-O, Cobia-Olmedo; D, Doebley; I, Iltis; K, Kato; P, Puga; PI, USDA Plant Introduction Station.

[b] This accession had small purple blotches on the kernels, an effect resembling that of *Pl-Blotched* allele of *pl1*.

[c] These tests produced only pale blue kernels, suggesting that these teosintes possess an additional modifer of *c2* activity.

substitutional differences. This observation suggests that the haplotypes can be divided into two distinct groups.

To test this hypothesis, sequences representing all 16 haplotypes were subjected to phylogenetic analysis. Bootstrap analysis provides strong (97%) support for partitioning the *c1* haplotypes into two discrete groups (Figure 3). The first group, which contains haplotypes 1–10, includes all the haplotypes isolated from maize and some haplotypes isolated from teosinte. We denote these the "class I" haplotypes. The other group, consisting of haplotypes 11–16, was found only in teosinte. We call these the "class II" haplotypes. The two classes of haplotypes can be discriminated by a number of molecular features. All class II haplotypes share both the deletion of the Box I gtgtc motif and a 4-bp deletion at sites −123 to −126 in Box II. The class II haplotypes also share many substitution polymorphisms relative to class I haplotypes, particularly in intron 1 (Figure 2).

Variation in the *c1* promoter: The region encompassing Boxes I and II is functionally distinct and important to gene expression (HATTORI *et al.* 1992; SCHEFFLER *et al.* 1994). Given the importance of this region, the amount of indel variation is surprising (Figure 2). It is reasonable to ask if there have been significantly more indel events in this short promoter region relative to other noncoding regions? To address this question, we must first make two assumptions: (1) each indel variant represents one and only one indel event and

Haplotype	5'UT				EX1	IN1
	333333333333333222211111	1	11111	1	1	1
	533333100000000999555444	3	22222	2	390	5
	364320087654321098741098	2I	76543	2I	464	6I
1	gccatgcttatattgaaaatgtgtc.....gtgca.....				tgc	....
2	----------.....-----.....				---	....
3	----------.....-----.....				---	....
4	a---------.....-----.....				---	....
5	----------.....-----.....				---	....
6	---g------.....-----.....				---	....
7	----------atgcac.....				---	....
8	----------.....-----.....				---	....
9	----------.....-----.....				---	....
10	----------.....-----gcc.....				---	....
11	--a-c----.....------.....				c--	ttaa
12	-ta---t.....-.....-.....agtttag				c--	....
13	--a-c----.....------.....				c--	....
14	--a-c----.....------.....				cat	....
15	--a-ca---.....-----c.....				c-t	....
16	--a-ca---.....-----c.....				c-t	....

Box I | Box II | SSS

Haplotype	IN1	EX2	IN2					EX3
	111111 122222	2	333334444444444	4	44	4	4444445	5
	557788 800112	7	789991222233333	5	56	7	7889991	5
	782448I925195	6	450120125601234	5I	77	2I	9890190	5
1	ctatg..tggccc	g	tatggggcattcgag.....tg....ttcattg					c
2	-----..c-----	-	-g----------.....--....-------					-
3	-----..------	-	----------.....--....c....c-					a
4	-----..------	-	----------.....--....c....c-					a
5	-----..------	-	----------.....--....c....c-					-
6	-----..------	-	----------.....--....-------					-
7	-----..------	-	----------.....--....c....c-					-
8	-----..------	-	g-...-ag..-----.....--....c....c-					-
9	-c---..------	-	g-...-ag..-----.....--....c....c-					-
10	-----..------	-	g-...-ag..-----.....--....c....c-					-
11	t-c--ta-a----	t	g-...----------.....c-....-------					-
12	t-cccta-a-ttt	-	g---------....attccgg-.catg-....-a					-
13	t-cccta-a-ttt	-	g---------....attccgg-.catg-....-a					-
14	t-c-cta-aat--	-	----------....a.....-.catg-....-a					-
15	t-c-cta-aat--	-	-----t---------.....--....-.....--					-
16	t-c-cta-aat--	-	----------.....--....c.....--					-

N N

FIGURE 2.—Nucleotide polymorphism data for *c1*. Haplotype numbers are given in Table 1. The reference numbers above nucleotide characters refer to the *c1* wild-type sequence of PAZ-ARES *et al.* (1987). Abbreviations are as follows: 5′ UT, 5′ untranslated regions; EX1, EX2 and EX3, exons 1, 2 and 3, respectively; IN1 and IN2, introns 1 and 2; S, a synonymous substitution within coding sequence; N, a nonsynonymous substitution. Box I and Box II refer to regions of the promoter. ---, sequence identity; . . ., an indel.

(2) each nucleotide site has an equal probability of being the start-point for an indel event under the null hypothesis that indels are equally distributed throughout noncoding regions. Using these assumptions, there are 17 total indel events in noncoding regions (Figure 2). Six of these indels are found in the promoter region

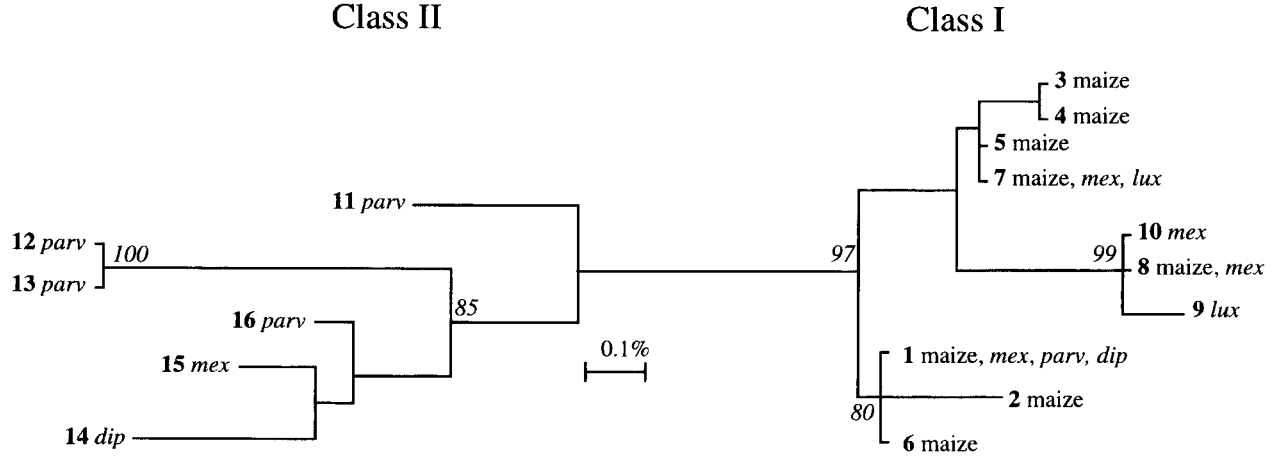

FIGURE 3.—An unrooted neighbor-joining phylogeny showing the relationships among haplotypes. Haplotype numbers are given in bold. The taxa of origin are given as maize, *mex* (*Z. mays* ssp. *mexicana*), *parv* (*Zea mays* ssp. *parviglumis*), *dip* (*Z. diploperennis*) and *lux* (*Z. luxurians*). Bootstrap values ≥80% are given in italics. The scale bar gives a rough indication of sequence divergence.

(which has a maximal length of 32 bp), and 11 are found in the remaining noncoding regions (which has a maximal length of 591 bp). If indels occur randomly throughout noncoding regions, the number of indels within the promoter region would be distributed as a binomial random variable with $n = 17$ and $p = 32/(32 + 591) = 0.051$ (as per LEICHT *et al.* 1995). Using this approach, the probability of observing at least five indels in the promoter region is 0.001, suggesting that indels are overrepresented in the *c1* promoter.

Recombination: Recombination can affect variation in the number of substitutions or indels among regions of a gene. For this reason, it is of interest to examine the role of recombination in diversification of *c1* haplotypes. We used the algorithm of HUDSON and KAPLAN (1985: Appendix 2) to estimate the minimum number of recombination events in a sample of genes. Using this method, there are no detectable recombination events in the sample of maize *c1* sequences. There are also no detectable events in the sample of five ssp. *mexicana* sequences. However, three recombination events are estimated to have occurred in the sample of the five *parviglumis* sequences. If one applies the algorithm to the entire sample of 27 sequences (which requires the assumption that all 16 haplotypes have had the opportunity to recombine), there is evidence for only three recombination events.

Recombination in maize *c1* sequences can be compared to recombination in other maize loci. In *Adh1,* recombination is clearly a potent force in generating haplotype diversity (GAUT and CLEGG 1993); the algorithm of HUDSON and KAPLAN (1985) detects a minimum of nine recombination events in a sample of six *Adh1* alleles, each of which is ≈2083 bp in length. Similarly, a minimum of five recombination events are found in a sample of 12 maize *Adh2* alleles (GOLOUBINOFF *et al.* 1993), each of which is only ≈331 bp in length. Thus, more recombination events are detectable in both *Adh1* and *Adh2* than in *c1,* despite a much smaller sample size in *Adh1* and much shorter sequences in *Adh2.* These results suggest either that recombination is more frequent in the two maize *Adh* loci than in *c1,* or that recombination events are difficult to detect in maize *c1* sequences because of relatively low levels of polymorphism (see below).

Nucleotide polymorphism and divergence among maize loci: Nucleotide diversity has been sampled at a number of maize loci. Figure 4 compares estimates of WATTERSON's (1975) θ and confidence intervals around $\hat{\theta}$ at seven maize loci. It has been shown that θ is heterogeneous among maize loci (SHATTUCK-EIDENS *et al.* 1990), suggesting that either substitution rates or population sizes (potentially a function of selection) differ among loci. Of these loci, only *Adh1, Adh2* and *c1* are known to encode functional proteins; the others represent anonymous single-copy regions of the maize nuclear genome (SHATTUCK-EIDENS *et al.* 1990). It should be mentioned that *Adh1, Adh2,* and *c1* were sampled over a broad geographic range; the anonymous loci were sampled only from U.S. inbred lines (SHATTUCK-EIDENS *et al.* 1990).

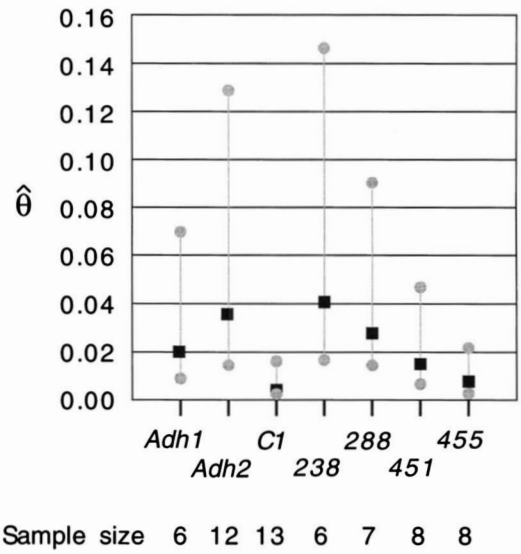

FIGURE 4.—Comparison of estimates of θ among maize loci. 95% confidence limits are given by shaded lines.

The *c1* locus is the least polymorphic locus that has been sampled in maize to date (Figure 4). Is the relative lack of variation in *c1* a function of neutral mutation rates, or does the lack of variation reflect a reduction of polymorphism due to selection? This question can be addressed using the HKA test to compare *r,* the ratio of intraspecific polymorphism to interspecific divergence, among loci. The HKA test investigates whether these ratios are heterogeneous among loci by comparing the fit of their estimates to an equilibrium neutral model (HUDSON *et al.* 1987). We applied the HKA test to *c1, Adh1,* and *Adh2* data, using sequences from either *Z. luxurians* or *Z. diploperennis* to measure divergence (Table 3). In neither case did tests reject the equilibrium neutral model at the 95% level. However, several features of the data suggest that HKA results must be interpreted with caution (see DISCUSSION). We also applied the test of TAJIMA (1989) to maize *c1* data. This test did not indicate significant deviation from an equilibrium neutral model ($D = -0.753$; $P > 0.05$).

Intertaxon comparisons of *c1* nucleotide polymorphism: Table 4 presents $\hat{\theta}$ for *c1* for three subspecies of *Z. mays.* The sample of sequences from maize contains less variation than those from ssp. *parviglumis* and ssp. *mexicana.* However, 95% confidence intervals for θ overlap, and thus we cannot conclude that θ is heterogeneous among these taxa at *c1.*

The three subspecies of *Z. mays* may have diverged within the past 75,000 years (APPENDIX), so that the gene pools of these taxa could be relatively homoge-

TABLE 3

Polymorphism and divergence in maize loci

Locus	*Z. diploperennis*			*Z. luxurians*		
	c1	*Adh1*	*Adh2*	*c1*	*Adh1*	*Adh2*
Polymorphism (%)[a]	0.37	2.13	3.47	0.37	2.13	3.47
Divergence (%)[b]	1.22	2.39	2.98	0.48	2.57	2.71
Ratio, r[c]	0.30	0.89	1.16	0.77	0.83	1.28
HKA[d]	0.104			0.062		

[a] Polymorphism was measured as average pairwise differences per nucleotide site between maize alleles based on sample sizes of 13, 6, and 12 sequences in *c1*, *Adh1*, and *Adh2*, respectively.

[b] Divergence was measured as average pairwise differences per nucleotide site between maize alleles and the alleles from the comparison taxa (*Z. luxurians* or *Z. diploperennis*) based on sample sizes of two sequences, one sequence, and two sequences per comparison taxa for *c1*, *Adh1*, and *Adh2*, respectively.

[c] Ratio of polymorphism within maize to the divergence between maize and the outgroup.

[d] Probabilities under HKA tests examining all three loci, based on 2000 simulations of the coalescent process.

neous. In addition, interspecific gene flow between taxa may retard the process of divergence. For these reasons, it is of interest to assess whether nucleotide sequence data at the *c1* locus provide any evidence of genetic subdivision among the three subspecies. Application of the method of HUDSON *et al.* (1992a) reveals evidence for subdivision between samples of maize and ssp. *parviglumis* sequences ($P = 0.000$) but no evidence for subdivision between maize and ssp. *mexicana* ($P = 0.390$). There is weak evidence for subdivision between ssp. *parviglumis* and ssp. *mexicana* ($P = 0.063$).

Introgression: All *Zea* taxa can hybridize, and isozyme studies provide evidence for introgression both between maize and *Z. luxurians* and between maize and *Z. diploperennis* (DOEBLEY *et al.* 1984). Thus, it is possible that the presence of a *c1* haplotype in more than one taxon is indicative of introgression. However, shared haplotypes can also reflect very recent divergences among taxa. We attempt to discriminate between these two hypotheses to determine whether introgression must be invoked to explain shared haplotypes.

We ask the question: Given rates of molecular evolution, divergence times between taxa and no introgression, does one expect to find zero differences between sequences from the different taxa? The only available estimate of the divergence time between maize and *mexicana* is ≈75,000 years (APPENDIX); this number reflects the minimum divergence time between sequences in the absence of introgression. Although little is known about absolute rates of nucleotide substitution in the *c1* locus, estimates of substitution rates in plant nuclear loci vary from $5-30 \times 10^{-9}$ synonymous substitutions per site per year (WOLFE *et al.* 1987; GAUT and CLEGG 1991). We make the conservative assumption that *c1* evolves at the low end of this range. Given the divergence time and the evolutionary rate, does one expect to find zero differences between maize and *mexicana* sequences in the absence of introgression, or must introgression (which reduces the divergence time) be invoked to explain the fact that identical sequences were found in two taxa?

We answer this question by simulating sequence evolution using the substitution model of KIMURA (1980) with the 4.7:1 transition:transversion ratio observed in our *c1* data (Figure 2). Given the evolutionary rate (5×10^{-9} substitutions per site per year) and a divergence time between hypothetical ssp. *mexicana* and maize sequences (75,000 years in the absence of introgression), we simulated the evolution of pairs of 960-bp sequences. Out of 1000 simulated sequence pairs, over 50% of the pairs had no differences between them. These simulations reveal that identical sequences in maize and ssp. *mexicana* are not unexpected, even in the absence of introgression, given the divergence time and the rate of nucleotide substitution. These results indicate that introgression need not be invoked to explain sequence identities among any of the subspecies of *Z. mays*.

This analysis can be applied to the sequence identity observed between *Z. mays* and species in the section *Luxuriantes*. The available estimate of divergence times from isozyme data suggests that the two sections of the genus Zea diverged ≈135,000 years ago (APPENDIX). With this divergence time, our simulations reveal that 20% of sequence pairs have no differences between them, suggesting both that identical haplotypes are not significantly rare (*i.e.*, expected in frequencies < 0.05), and introgression need not be invoked to explain

TABLE 4

Comparisons of $\hat{\theta}$ for *c1* in *Zea*

Taxa	m^a	n^b	$\hat{\theta}$	$\hat{\theta}_{0.025}$	$\hat{\theta}_{0.975}$
ssp. *mays*	709	13	0.004	0.002	0.014
ssp. *mexicana*	694	5	0.012	0.005	0.055
ssp. *parviglumis*	684	5	0.015	0.006	0.069

[a] Number of silent nucleotide sites compared.

[b] Number of alleles in the sample.

shared haplotypes among taxa from different sections of Zea.

A number of points must be made about these analyses. First, these analyses do not reject the hypothesis that introgression has occurred but simply suggest that introgression need not be invoked to explain the fact that identical sequences were found in different taxa. Second, divergence time estimates may be inaccurate. Third, these analyses are heavily assumption laden. Ideally, one would use a coalescent approach, which accounts for correlations due to phylogenetic relationships among sequences and considers haplotype frequencies, to examine whether multi-sequence profiles are consistent with a lack of introgression. However, given that there is evidence for selection at the *c1* locus (DISCUSSION), an equilibrium neutral coalescent model is not appropriate, and it is therefore difficult to select an appropriate coalescent model for study. Finally, because we apply the relatively rapid synonymous rate to the entire 960 bp when in fact much of the gene evolves at the much slower nonsynonymous rate, the likelihood of finding identical sequences in the different taxa should be even greater than our test indicates.

DISCUSSION

Allelic diversity at *c1*: Teosinte possesses dominant functional alleles at high frequency at all enzymatic loci that we assayed (Table 2). In contrast, at the two regulatory loci (*c1* and *r*), recessive alleles incapable of activating the anthocyanin pathway during kernel maturation predominate. At *r*, the dominant function allele for kernel color was absent; however, because we sampled only 45 plants, there is a clear possibility that functional *r* alleles exist in teosinte at low frequency. At *c1*, dominant *C1* was observed in a single teosinte population. This may indicate either that *C1* is native to teosinte or it may represent a recent introgression event from maize.

Our observations on allelic diversity at the anthocyanin loci in teosinte enable us to make several inferences about the evolution of the purple kernel phenotype. First, the evolution of this trait in maize must have been accomplished by changes at the regulatory loci rather than the enzymatic loci. This conforms to predictions that regulatory loci are key players in the evolution of new phenotypes (GOODRICH *et al.* 1992; DOEBLEY 1993). Second, the evolution of purple kernels required selection for functional alleles at both *c1* and *r*. This requirement raises the intriguing possibility that *C1* and *R* are harbored in different teosinte populations and that the evolution of purple kernels was accomplished when maize cultivators recombined existing allelic variation, creating a population that possessed functional alleles at both loci. Third, once functional *C1* and *R* alleles were combined, a variety of colors could be produced by combinations with other teosinte alleles such as *pr* to produce red kernels.

The nature of the changes in *r* and *c1* involved in the evolution of purple kernels can be inferred from our genetic tests. For *c1*, teosinte possesses the *c1-p* allele at high frequency, indicating that many teosintes make functional C1 protein, but only during germination and not during kernel maturation (CHEN and COE 1977). Similarly, teosinte possesses the functional plant component (*r-r* allele) at *r* indicating that teosinte makes functional R protein, but only in vegetative tissues and not in the kernel. These observations suggest that purple kernels evolved not by changes in the protein products of these genes since teosinte makes both functional proteins, but rather by changes in their *cis*-regulatory elements that enabled their activation during seed maturation. These inferences are consistent with much prior research that has shown that the different tissue specific patterns of anthocyanin expression result from *cis*-regulatory differences among alleles at regulatory genes (LUDWIG *et al.* 1990; HATTORI *et al.* 1992; RADICELLA *et al.* 1992; PATTERSON *et al.* 1995).

Introgression: Introgression can substantially impact the amount of polymorphism at a locus (RIESEBERG and WENDEL 1993). The issue of introgression arises for *c1* because haplotypes 1, 7 and 8 were found in more than one taxon and all species of *Zea* are interfertile. We take a new approach to investigate whether introgression or retention of an ancient haplotype explains this distribution. We calculated the probability that the haplotypes could have survived since divergence from the common ancestor without accumulating any mutational differences. Our simulation results indicate that the divergence times are sufficiently small for *c1* haplotypes to have been retained without alteration since the divergence from the common ancestor. Thus, introgression need not be invoked to explain the shared haplotypes. Our analysis does not preclude the possibility that introgression has occurred, but our approach to this question provides an objective criterion by which introgression and ancient retention can be distinguished. Given the historical interest in introgression and the growth of molecular sequence data for Zea and other genera, the probability that the haplotypes could have survived unaltered since divergence should be considered before introgression is inferred.

The case for reduction in polymorphism due to selection: While both class I and class II haplotypes were observed in teosinte, our maize sample contains only class I. Does this skewed distribution of haplotypes reflect the effects of past selection to eliminate class II haplotypes from the maize gene pool? Assume that haplotypic variation is neutral and partitioned into the maize and teosinte gene pools such that class I and class II variants enter the gene pools randomly from the

common ancestor. This simple model, in which allelic types sort randomly into gene pools, is consistent with previous findings that suggested that the common ancestor to *Zea* species was highly polymorphic with lineage sorting partitioning variation among taxa (GAUT and CLEGG 1993; GOLOUBINOFF *et al.* 1993). Class II haplotypes are found at a frequency of 0.58, with a 95% confidence interval of 0.30–0.88 (using the normal approximation), in our teosinte samples. If the process of lineage sorting is roughly equivalent among taxa, the probability of seeing no class II haplotypes in a maize sample of 13 alleles is quite low [P (no class II haplotypes) = $(1 - 0.58)^{13} < 0.001$]. This result holds even when the lower bound estimate of class II frequency is used [P (no class II haplotypes) = $(1 - 0.30)^{13} = 0.01$]. Although the "lineage sorting" model on which these estimates are based is quite simplistic, these calculations tend to suggest that class II haplotypes are underrepresented in the maize gene pool. One explanation for this underrepresentation is that selection has acted to remove class II haplotypes from the maize gene pool.

It should be noted that two class II alleles (the *c1-n* and *c1-p* alleles) have been isolated from maize. These two alleles were isolated because of their mutant phenotypes, the lack of anthocyanin pigmentation in kernels (SCHEFFLER *et al.* 1994). For this reason, they were not included in our random sample of maize alleles. However, inclusion of these haplotypes in the above analysis does not dramatically alter the conclusion that class II haplotypes are underrepresented in the maize sample (data not shown).

If selection has acted to reduce polymorphism in the *c1* locus of maize, it is important to consider its nature. Our sample of maize alleles contains only class I haplotypes, but these class I haplotypes likely predate the origin of maize. Among the class I haplotypes found within maize, the maximal number of nucleotide differences between sequence pairs is seven nucleotides. If *c1* evolves within the range stated for plant nuclear genes (WOLFE *et al.* 1987), then the two most different haplotypes found in maize diverged ≈140,000–850,000 years ago. While these estimates have large variances, they suggest that the oldest class I haplotypes found within maize diverged before the domestication of maize ≈7500 years ago (ILTIS 1983).

The age of class I haplotypes in maize suggests that reduced polymorphism at the *c1* locus is not the result of a selective sweep in maize, where a new mutant has arisen and swept to fixation in the maize lineage. This view is bolstered by the nonsignificant TAJIMA (1989) test, which has reasonable power to detect a sweep due to hitchhiking selection (BRAVERMAN *et al.* 1995). Rather, it suggests that selection has acted to reduce the frequency of class II haplotypes in maize. Selection could be on *c1* itself or on a linked locus. If selection is on the *c1* locus, it is reasonable to hypothesize that anthocyanin pigmentation in maturing kernels has been selected during or after domestication.

HKA tests: Several lines of evidence suggest that polymorphism at *c1* in maize has been reduced as a result of selection, but HKA tests do not reject the null hypothesis of equilibrium neutral evolution (Table 3). A number of features of the HKA test, as applied here, suggest that a lack of rejection of the null hypothesis should not be viewed as evidence to support neutrality.

First, the HKA test probably lacks power with these data. Lineage sorting has played a substantial role in partitioning variation among *Zea* taxa, so that the depth of allelic lineages within a taxon can be far greater than the time of divergence between taxa. In this case, both polymorphism and divergence measures have large variances, resulting in low statistical power to reject the null hypothesis. Second, the model used in the HKA test does not include introgression, and thus by applying this model, we have implicitly assumed that introgression does not occur (or occurs at very low levels) between maize and the comparison taxa (*Z. diploperennis* and *Z. luxurians*). This is a questionable assumption, because introgression has been reported between maize and both comparison taxa (DOEBLEY *et al.* 1984). The effect of introgression on HKA tests is not clear, as the effect will vary with the magnitude and direction of introgression.

Systematics of *Zea*: Analysis of isozymes and chromosomal knobs suggest that ssp. *parviglumis* is the closest relative to maize (APPENDIX). However, subdivision tests suggest that maize and ssp. *mexicana* are closely related at the *c1* locus, while maize and ssp. *parviglumis* are genetically subdivided. Are these observations inconsistent with the hypothesis that ssp. *parviglumis* is the closest relative to maize? The answer is no. First, it is apparent that lineage sorting is an important evolutionary process in *Zea*, such that gene trees may not reflect species' relationships. The inference of relationships in a lineage sorting system is further clouded by selection because selection on *c1* appears to have biased the partitioning of alleles into gene pools. Second, the possibility of interspecific gene flow confounds the inference of relationship by phylogenetic descent. Assume that maize and ssp. *parviglumis* are closest relatives. If introgression occurs frequently between maize and ssp. *mexicana* and infrequently between maize and ssp. *parviglumis*, then molecular phylogenies will not reflect the phylogenetic "truth" of a maize-ssp. *parviglumis* clade. This could be the case with our *c1* data since ssp. *mexicana* hybridizes frequently with maize while ssp. *parviglumis* does not (WILKES 1977; DOEBLEY 1990b).

Molecular and phenotypic evolution: One model for the evolution of purple kernels in maize could be that new mutation at *c1* in maize after its divergence from teosinte led to the activation of anthocyanin synthesis during kernel maturation. Both our sequence data and

allelic survey suggest that this was not the case. First, we demonstrate that dominant *C1* exists in teosinte, albeit as a rare allele (Table 2). Second, many teosinte *c1* alleles produce a functional *c1* protein as indicated by the presence of *c1-p* at a frequency of 0.51, and eight of 14 teosinte sequence samples possessed the gtgtc motif in Box I required for *c1* expression during kernel maturation (Figure 2). Thus, teosinte contains the necessary components of *c1* to activate anthocyanin synthesis in the kernel. If new mutation was not involved, how did colored kernels evolve?

The simplest explanation would be that both class I (including *C1*) and class II haplotypes existed in teosinte before the origin of maize. Then, during or after the domestication of maize a preference for colored kernels by ancient agriculturalists caused selective elimination of class II haplotypes, which lack the Box I gtgtc motif, and a corresponding increase in the frequency of class I haplotypes, which contain this motif and thus can activate the pathway during kernel maturation. Paradoxically under this scenario, a reduction in haplotype diversity would be associated with an increase in phenotypic diversity since the presence of *C1* (plus *R*) in maize would uncover variation at other anthocyanin loci (Table 2), producing different shades of purple, blue and red. In teosinte, with only *c1* (or *C1-I*) and *r*, variation at these other loci should have no visible effect of kernel color. This scenario is consistent with both our sequence data and allelic survey.

Another possibility is that intragenic recombination at *c1* has played a role in the evolution of colored kernels. Consider three observations. (1) Eight of the 14 teosinte *c1* samples possess the gtgtc motif in Box I of the promoter that is required for the production of anthocyanin during kernel maturation (Figure 2). (2) Fifty-one percent of the recessive *c1* alleles observed in the allelic survey were of the *c1-p* class, indicating that they encode a functional protein. (3) Only one teosinte population from the allelic survey possessed the dominant *C1* (Table 2) despite the fact the necessary promoter and protein moieties are relatively common in teosinte. These observations can be reconciled if teosinte alleles fall largely into three groups: those with functional promoters (*i.e.*, with the gtgtc motif) and disfunctional proteins (class I haplotypes and *C1-I* or *c1-n* alleles), those with nonfunctional promoters and functional proteins (class II haplotypes and *c1-p* alleles), and those with nonfunctional promoters and disfunctional proteins (class II haplotypes and *c1-n* alleles). If this is the case, then dominant *C1* could have evolved by a recombination event that combined the gtgtc motif with a functional protein coding region. Since our sequence data do not cover the full coding region of the gene, they can not confirm nor refute this possibility.

Finally, there is an apparent conflict between our sequence data and allelic survey since three of 14 teosinte samples had sequences identical to the functional *C1* allele of maize, although functional *C1* alleles were rare in the allelic survey. There are at least two possible explanations for this apparent conflict. First, since our sequence data do not cover the full coding region, it is possible that, although some teosinte alleles are identical to maize *C1* for the part of *c1* we sequenced, they possess lesions in 3′ regions we have not sequenced, rendering them incapable of activating anthocyanin synthesis. Second, the *C1*-like teosinte alleles may possess distant 5′ or 3′ regulatory sequences or methylation patterns that render them inactive. Phenomena of this nature have been demonstrated or inferred for *b* and *pl* (COCCIOLONE and CONE 1993; PATTERSON *et al.* 1995).

The authors thank Dr. KAREN CONE for the *c1* clone, Dr. JODY HEY for the use of his coalescent simulation program, Dr. DON MCCARTY for the 5′ primer, and Dr. VIRGINIA WALBOT for helpful suggestions on the genetic tests. This work supported by grants from the Graduate School of the University of Minnesota, National Science Foundation (BSR-9107175), Minnesota Agricultural Experiment Station, and U.S. Department of Agriculture (95-00566 to B.S.G.).

LITERATURE CITED

BRAVERMAN, J. M., R. R. HUDSON, N. L. KAPLAN, C. H. LANGLEY and W. STEPHAN, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140:** 783–796.

CHEN, S.-M., and COE, E. H., 1977 Control of anthocyanin synthesis by the C locus in maize. Biochem. Genet. **15:** 333–346.

COCCIOLONE, S. M., and K. C. CONE, 1993 *Pl-Bh*, an anthocyanin regulatory gene of maize that leads to variegated pigmentation. Genetics **135:** 575–588.

COE, E. H., D. A. HOISINGTON and M. G. NEUFFER, 1988 The genetics of corn, pp. 81–258 in *Corn and Corn Improvement*, edited by G. F. SPRAGUE and J. W. DUDLEY. The American Society of Agronomy, Madison, WI.

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

DOEBLEY, J., 1990b Molecular evidence for gene flow among *Zea* species. Bioscience **40:** 443–448.

DOEBLEY, J., 1993 Genetics, development and plant evolution. Curr. Opin. Gen. Dev. **3:** 865–872.

DOEBLEY, J., M. M. GOODMAN and C. W. STUBER, 1984 Isoenzymatic variation in *Zea* (Gramineae). Syst. Bot. **9:** 203–218.

DOEBLEY, J. F., W. RENFROE and A. BLANTON, 1987 Restriction site variation in the *Zea* chloroplast genome. Genetics **117:** 139–147.

DOONER, H. K., T. P. ROBBINS and R. A. JORGENSEN, 1991 Genetic and developmental control of anthocyanin biosynthesis. Annu. Rev. Genet. **25:** 173–199.

FELSENSTEIN, J., 1993 PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

GAUT, B. S., and M. T. CLEGG, 1991 Molecular evolution of *alcohol dehydrogenase 1* in members of the grass family. Proc. Natl. Acad. Sci. USA **88:** 2060–2064.

GAUT, B. S., and M. T. CLEGG, 1993 Molecular evolution of the *Adh1* locus in the genus *Zea*. Proc. Natl. Acad. Sci. USA **90:** 5095–5099.

GOFF, S. A., K. C. CONE and M. E. FROMM, 1991 Identification of functional domains in the maize activator *C1:* comparison of wild-type and dominant inhibitor proteins. Genes Dev. **5:** 298–309.

GOLOUBINOFF, P., S. PAABO and A. C. WILSON, 1993 Evolution of maize inferred from sequence diversity of an *Adh2* gene segment from archaeological specimens. Proc. Natl. Acad. Sci. USA **90:** 1997–2001.

GOODMAN, M. M., 1978 A brief survey of the races of maize and current attempts to infer racial relationships, pp. 143–158 in

Maize Breeding and Genetics, edited by D. B. WALDEN. John Wiley and Sons, New York.

GOODRICH, J., R. CARPENTEUR and E. COEN, 1992 A common gene regulates pigmentation pattern in diverse plant species. Cell **68:** 955–964.

HATTORI, T., V. VASIL, L. ROSENKRANS, L. C. HANNAH, D. R. MCCARTY *et al.*, 1992 The *Viviparous-1* gene and abscisic acid activate the *C1* regulatory gene for anthocyanin biosynthesis during seed maturation in maize. Genes Dev. **6:** 609–618.

HILTON, H., R. M. KLIMAN and J. HEY, 1994 Using hitchhiking genes to study adaptation and divergence during speciation within the *Drosophila melanogaster* species complex. Evolution **48:** 1990–1913.

HOLTON, T. A., and E. C. CORNISH, 1995 Genetics and biochemistry of anthocyanin biosynthesis. Plant Cell **7:** 1071–1083.

HUDSON, R. R., and N. L. KAPLAN, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

HUDSON, R. R., M. KREITMAN and M. AGUADE, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

HUDSON, R. R., D. D. BOOS and N. L. KAPLAN, 1992a A statistical test for detecting geographic subdivision. Mol. Biol. Evol. **9:** 138–151.

HUDSON, R. R., M. SLATKIN and W. P. MADDISON, 1992b Estimation of levels of gene flow from DNA sequence data. Genetics **132:** 583–589.

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

ILTIS, H. H., and J. F. DOEBLEY, 1980 Taxonomy of *Zea* (Gramineae). II. Subspecific categories in the *Zea mays* complex and a generic synopsis. Amer. J. Bot. **67:** 994–1004.

KATO, T. A., 1976 *Cytological Studies of Maize.* Mass. Agric. Exper. Station Res. Bull. No. 635.

KIMURA, M., 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16:** 111–120.

KREITMAN, M., and R. R. HUDSON, 1991 Inferring the evolutionary histories of *Adh* and the *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics **127:** 565–582.

LEICHT, B. G., S. V. MUSE, M. HANCZYC and A. G. CLARK, 1995 Constraints on intron evolution in the gene encoding the myosin alkali light chain in *Drosophila.* Genetics **139:** 299–308.

LUDWIG, S., B. BOWEN, L. BEACH and S. WESSLER, 1990 A regulatory gene as a novel visible marker for maize transformation. Science **247:** 449–450.

MCCARTY, D. R., C. B. CARSON, P. S. STINARD and D. S. ROBERTSON, 1989 Molecular analysis of *viviparous-1:* an abscisic acid-insensitive mutant of maize. Plant Cell **1:** 523–532.

MCCLINTOCK, B., T. A. KATO and A. BLUMENSCHEIN, 1981 *Chromosome Constitution of Races of Maize.* Colegio de Postgraduados, Chapingo, Mexico.

NEI, M., 1975 *Molecular Population Genetics and Evolution.* North-Holland Co., Amsterdam.

PATTERSON, G. I., K. M. KUBO, T. SHROYER and V. L. CHANDLER, 1995 Sequences required for paramutation of the maize *b* gene map to a region containing the promoter and upstream sequences. Genetics **140:** 1389–1406.

PAZ-ARES, J., D. GHOSAL, U. WIENAND, P. A. PETERSON and H. SAEDLER, 1987 The regulatory *C1* locus of *Zea mays* encodes a protein with homology to *myb* proto-oncogene products and with structural similarities to transcriptional activators. EMBO J. **6:** 3553–3558.

PAZ-ARES, J., D. GHOSAL and H. SAEDLER, 1990 Molecular analysis of the *C1-I* allele from *Zea mays:* a dominant mutant of the regulatory *C1* locus. EMBO J. **9:** 315–321.

RADICELLA, J. P., D. BROWN, L. A. TOLAR and V. L. CHANDLER, 1992 Allelic diversity of the maize *B* regulatory gene: different leader and promoter sequences of two *B* alleles determine distinct tissue specificities of anthocyanin production. Genes Dev. **6:** 2152–2164.

RIESEBERG, L. H., and J. F. WENDEL, 1993 Introgression and its consequences in plants, pp. 70–109 in *Hybrid Zones and the Evolutionary Process,* edited by R. HARRISON. Oxford University Press, Oxford.

SAITOU, N., and M. NEI, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

SCHEFFLER, B., P. FRANKEN, A. SCHRELL, H. SAEDLER and U. WIENAND, 1994 Molecular analysis of *C1* alleles in *Zea mays* defines regions involved in the expression of this regulatory gene. Mol. Gen. Genet. **242:** 40–48.

SHATTUCK-EIDENS, D. M., R. N. BELL, S. L. NEUHAUSEN and T. HELENTJARIS, 1990 DNA sequence variation within maize and melon: observations from polymerase chain reaction amplification and direct sequencing. Genetics **126:** 207–217.

SMITH, B. D., 1995 *The Emergence of Agriculture.* W. H. Freeman, New York.

SWOFFORD, D. L., and D. P. BEGLE, 1993 PAUP: phylogenetic analysis using parsimony. Illinois Natural History Survey, Champaign, IL.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Pop. Biol. **7:** 188–193.

WILKES, G., 1967 *Teosinte: The Closest Relative of Maize.* Bussey Institute, Harvard University, Cambridge, MA.

WILKES, G., 1977 Hybridization of maize and teosinte, in Mexico and Guatemala and the improvement of maize. Econ. Bot. **31:** 254–293.

WILSON, A. C., 1976 Gene regulation in evolution, pp. 225–234 in *Molecular Evolution,* edited by F. J. AYALA. Sinauer Associates, Sunderland, MA.

WOLFE, K. H., W.-H. LI and P. M. SHARP, 1987 Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast and nuclear DNAs. Proc. Natl. Acad. Sci. USA **84:** 9054–9058.

ZIMMER, E. A., E. R. JUPE and V. WALBOT, 1988 Ribosomal gene structure, variation and inheritance in maize and its ancestors. Genetics **120:** 1125–1136.

Communicating editor: A. G. CLARK

APPENDIX

Phylogeny of *Zea:* DOEBLEY (1990a) summarized evidence from isozymes and chloroplast DNA for Zea phylogeny. To test the reliability of the cpDNA phylogeny, we performed parsimony analysis with 100 bootstrap replications using PAUP version 3.1 (SWOFFORD and BEGLE 1993). Both sections *Luxuriantes* and *Zea* were monophyletic in all 100 replications, providing strong evidence that the chloroplast genomes of the taxa of the two sections are monophyletic (Figure 5a). All subsectional clades in the cpDNA tree lack strong statistical support (*i.e.*, bootstrap values below 90/100).

To clarify the relationships among the subspecies of *Z. mays*, we constructed a neighbor-joining tree using the modified Rogers' distances for isozymes presented by DOEBLEY *et al.* (1984). We also analyzed the chromosome knob data of KATO (1976) and MCCLINTOCK *et al.* (1981). The latter data included all teosinte accessions from KATO (1976) and a sample of 61 maize accessions from throughout Mexico, Central America and the U.S. from MCCLINTOCK *et al.* (1981). The analysis included all knob positions scored by the authors. For each knob position, an index was constructed by multiplying the number of large, medium and small knobs by 3, 2 and 1, respectively. The index was standardized such that

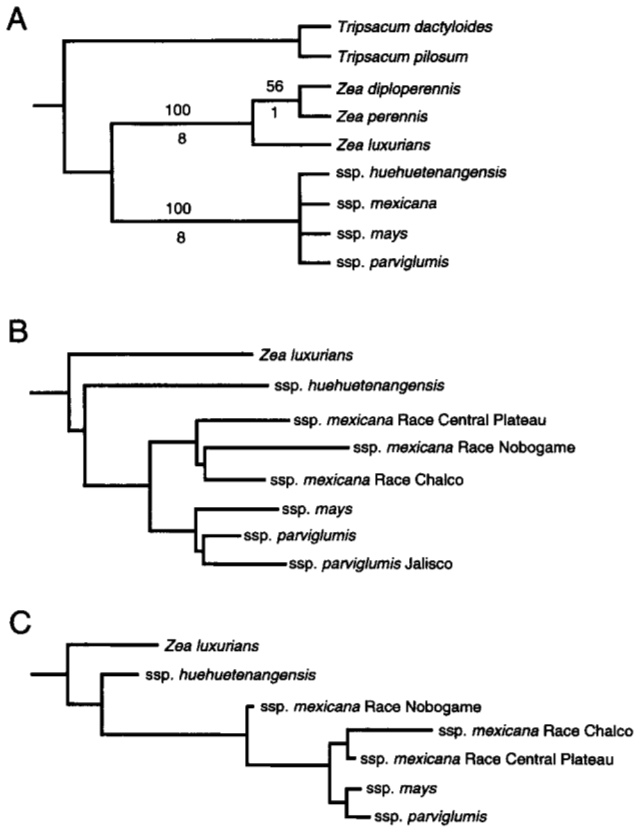

FIGURE 5.—Phylogenetic trees for *Zea*. (A) Parsimony tree based on the cpDNA data from DOEBLEY (1990a). The numbers above the branches indicate the number of the 100 bootstrap samples in which the clade was observed; the numbers below the branches indicate the number of steps on that branch. The tree was rooted with *Tripsacum dactyloides* and *T. pilosum*. (B) Neighbor-joining tree based on isozyme data from DOEBLEY *et al.* (1984). This tree used modified Rogers' distances and was rooted with *Z. luxurians*. (C) A continuous character maximum likelihood tree based on chromosome knob data from KATO (1976) and MCCLINTOCK *et al.* (1981). The tree was rooted with *Z. luxurians*.

an accession with all large knobs would have a value of 1.0, and one that was completely knobless would have a value of 0.0. These data were then analyzed by the CONTML version 3.57 program of the PHYLIP package (FELSENSTEIN 1993). Both isozyme and knob data indicate that maize shares its closest relationship to ssp. *parviglumis* and a more distant relationship to ssp. *mexicana* (Figure 5, B and C).

TABLE 5

Divergence times for *Zea*

Taxa	Divergence time (B.P.)[a]
ssp. *mexicana*-ssp. *parviglumis*	61,000
ssp. *mexicana*-ssp *mays*	75,500
ssp. *parviglumis*-ssp. *mays*	18,500
ssp. *parviglumis* (Balsas)-ssp. *mays*	12,659
Zea diploperennis-Zea mays	134,500
Zea luxurians-Zea mays	135,167

[a] Values are presented in years before present (B.P.) and were calculated with an α of 10^{-6}.

Divergence times for *Zea* taxa: Isozyme data can be used to estimate divergence times among taxa with the formula $t = D/2\alpha$, where t is time, D is genetic distance, and α is the mutation rate (NEI 1975). Mutation rates for isozymes are estimated to vary from 10^{-5} to 10^{-7} (NEI 1975). We apply this method of estimating divergence times to isozyme data for *Zea* (DOEBLEY *et al.* 1984). The results can vary widely depending upon α; however, there are independent criteria for selecting the appropriate value for α. (1) The archaeological record establishes that maize existed (*i.e.*, had diverged from teosinte) by 5000 B.P. (SMITH 1995), thus values for α that yield divergence times for maize-ssp. *parviglumis* smaller than this are clearly inaccurate. (2) Agriculture did not exist in the New World before 15,000 years ago (SMITH 1995), thus values for α that yield dates larger than this are not credible. Applying these criteria, 10^{-6} appears to be the appropriate α, since 10^{-7} would place the maize-ssp. *parviglumis* divergence at 185,000 B.P. and 10^{-5} would place it at 1850. Table 5 presents divergence times for *Zea* taxa. We emphasize that these estimates are based on numerous assumptions, such as a mutation rate that is homogeneous over time and across taxa, that can not be verified. Nevertheless, they provide reasonable first estimates of the divergence times for *Zea* taxa.