DOI: 10.1534/genetics.105.053942

Gene Conversion Between Direct Noncoding Repeats Promotes Genetic and Phenotypic Diversity at a Regulatory Locus of *Zea mays* (L.)

Feng Zhang[1] and Thomas Peterson[2]

Department of Genetics, Development and Cell Biology and Department of Agronomy, Iowa State University, Ames, Iowa 50011

Manuscript received November 25, 2005
Accepted for publication June 19, 2006

ABSTRACT

While evolution of coding sequences has been intensively studied, diversification of noncoding regulatory regions remains poorly understood. In this study, we investigated the molecular evolution of an enhancer region located 5 kb upstream of the transcription start site of the maize *pericarp color1* (*p1*) gene. The *p1* gene encodes an R2R3 Myb-like transcription factor that regulates the flavonoid biosynthetic pathway in maize floral organs. Distinct *p1* alleles exhibit organ-specific expression patterns on kernel pericarp and cob glumes. A cob glume-specific regulatory region has been identified in the distal enhancer. Further characterization of 6 single-copy *p1* alleles, including *P1-rr* (red pericarp/red cob) and *P1-rw* (red pericarp and white cob), reveals 3 distinct enhancer types. Sequence variations in the enhancer are correlated with the *p1* gene expression patterns in cob glume. Structural comparisons and phylogenetic analyses suggest that evolution of the enhancer region is likely driven by gene conversion between long direct noncoding repeats (~6 kb in length). Given that tandem and segmental duplications are common in both animal and plant genomes, our studies suggest that recombination between noncoding duplicated sequences could play an important role in creating genetic and phenotypic variations.

EVOLUTION of *cis*-regulatory elements has been indicated as a major contributor to phenotypic variation, because changes in regulatory regions can induce temporal and spatial expression pattern changes (Doebley and Lukens 1998; Ludwig 2002; Wray *et al.* 2003; Carroll *et al.* 2004). Despite the importance of *cis*-elements in regulation of gene expression, the molecular basis of the *cis*-regulatory region evolution remains poorly understood. In recent studies, comparisons of *cis*-regulatory regions between natural variants have been applied to investigate their molecular evolution (Hanson *et al.* 1996; Wang *et al.* 1999; Ludwig *et al.* 2000; Purugganan 2000). In this study, we used natural variations of the gene specifying flavonoid pigment patterns in maize to gain insight into the evolutionary dynamic of *cis*-regulatory sequences. Several genetic loci, including the *c1* (*colorless1*), *p1* (*pericarp color1*), *r1* (*red1*), *b1* (*booster1*), and *pl1* (*purple plant1*) genes, that regulate flavonoid biosynthetic pathways of maize have been well characterized at the molecular level. All of these regulatory genes display diverse patterns of gene expression in both floral and vegetative organs of maize (Ludwig and Wessler 1990; Cone *et al.* 1993; Procissi *et al.* 1997; Selinger *et al.* 1998). Allelic variation has been investigated at the *c1*, *b1*, and *r1* loci in detail. Variations in regulatory regions have been linked to distinct allelic expression patterns and phenotypic diversity (Selinger *et al.* 1998; Li *et al.* 2001). Moreover, studies on the *c1* gene have suggested that the *cis*-regulatory region could be subject to selection, resulting in increased frequency of one *c1* haplotype and giving rise to the pigmented kernel aleurone phenotype (Hanson *et al.* 1996).

Recent analyses of the maize *p1* gene provide further evidence that variations in *cis*-regulatory regions are involved in phenotypic diversification. The *p1* gene encodes an R2R3 Myb-like transcription factor (Jiang *et al.* 2004). Expression of the *p1* gene is observed predominantly in the floral organs, most notably the kernel pericarp and cob glumes. According to the pigmentation patterns in these two organs, *p1* alleles are commonly classified into four major types: *P1-rr* (red pericarp/red cob), *P1-wr* (white pericarp/red cob), *P1-rw* (red pericarp/white cob) and *p1-ww* (white pericarp/white cob). Three distinct *p1* alleles, *P1-rr4B2*, *P1-rw1077*, and *P1-wr[w22]*, have been characterized and compared at the molecular level. All of these *p1* alleles share highly similar coding regions, but differ in the flanking regulatory sequences (Chopra *et al.* 1998; Sidorenko *et al.* 2000). Our recent data have demonstrated that changes in the enhancer region, located 5 kb 5′ of the transcription start site, result in phenotypic variations between *P1-rr4B2* and *P1-rw1077* (Zhang and Peterson 2005).

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. DQ160218–DQ160223.

[1]*Present address:* Department of Plant Biology, University of Georgia, Athens, GA 30603.

[2]*Corresponding author:* 2208 Molecular Biology Bldg., Iowa State University, Ames, IA 50011. E-mail: thomasp@iastate.edu

TABLE 1

Collections of single-copy *p1* alleles

Allelic name	Allelic type	Phenotype		Source
		Pericarp	Cob glume	
P1-rr4B2	*P1-rr4B2*	Red	Red	Derivative from *P1-vv*[a]
P1-rrCFS181	*P1-rr4B2*	Red	Red	Brink's collection[b]
P1-rr13:255 A10	*P1-rr4B2*	Red	Red	From *P1-vv*
P1-rr1088	*P1-rr1088*	Red	Red	To be determined
P1-rrCFS36	*P1-rrCFS36*	Red	Red	Brink's collection
P1-rrCFS33	*P1-rrCFS36*	Red	Red	Brink's collection
P1-rrCFS305	*P1-rrCFS36*	Red	Red	Brink's collection
P1-rrCFS548	*P1-rrCFS36*	Red	Red	Brink's collection
P1-rrCFS272	*P1-rrCFS36*	Red	Red	Brink's collection
P1-rw1077	*P1-rw1077*	Red	White	Maize Genetic Coop
P1-rwCFS325	*P1-rw1077*	Red	White	Brink's collection
P1-rwCFS302	*P1-rwCFS302*	Red	White	Brink's collection
P1-rwCFS332	*P1-rwCFS302*	Red	White	Brink's collection
P1-rwCFS342	*P1-rwCFS342*	Red	White	Brink's collection
P1-rwCFS334	*P1-rwCFS342*	Red	White	Brink's collection

[a] Lechelt *et al.* (1989).
[b] Brink and Styles (1966).

More than 100 natural variants of *p1* have been reported, each exhibiting distinctive patterns and intensity of pigmentation in kernel pericarp and cob glumes (Brink and Styles 1966). This rich collection of naturally occurring alleles provides an excellent system to study the evolution of *cis*-regulatory regions at the *p1* locus. In this study, 6 distinct single-copy *p1* alleles were characterized and compared. The goals of the study were: (i) to investigate DNA variations in noncoding regions of the *p1* locus, particularly the distal enhancer regions; (ii) to examine correlation between variations in *cis*-regulatory regions and phenotypic changes; and (iii) to identify the forces affecting evolution of the regulatory regions and the generation of genetic and phenotypic diversity at the *p1* locus.

MATERIALS AND METHODS

Maize germplasm: The *p1* alleles used in this study are homozygous in the 4Co63 inbred genetic background. As shown in Table 1, *p1* alleles with an assigned CFS number were from Brink's *p1* collection (Brink and Styles 1966). The *P1-rr4B2* and *P1-rw1077* alleles are the same as used in previous studies (Lechelt *et al.* 1989; Zhang and Peterson 2005).

DNA gel blot analysis: Genomic DNA was extracted from maize seedling leaves by the CTAB method (Saghai-Maroof *et al.* 1984) and digested with restriction enzymes according to manufacturer's instructions. Gel electrophoresis and hybridization procedures were conducted as described in previous studies (Sidorenko *et al.* 2000). The blot was probed with the *p1*-specific probe, genomic fragment 15 (Figures 1B and 2). *p1* alleles that exhibited identical hybridization patterns following digestion with *Sal*I, *Sac*I, *Eco*RI, and *Xba*I and probing with fragment 15 were grouped as the same allelic type.

Amplification and sequencing of the *p1* noncoding regions: Nested genomic PCR was performed to amplify ~6 kb from the 5′ noncoding regions. The primer pair, P1rr-18 and PA-B4, was used in the first-round reaction. A 1-μl aliquot of the first-round PCR product was subjected to a second-round PCR with the primer pair, P1rr-14 and PA-B6. For the *p1* allele, *P1-rrCFS36*, two additional primers, P1rr-32 and P1rr-25, were used due to the presence of a 1.6-kb transposon-like sequence inserted in the 5′ copy of fragment 15 (Figure 2). The 3′ noncoding regions were amplified in two overlapping pieces separately by using the primers EP5-16 and P1rr-16, as well as nested primers, P1rr-13r and P1rr-30 and P1rr-16r and P1rr-29. The 723-bp sequences in the second intron were amplified with primers 723-5 and 723-3. Locations and sequences of all primers are shown in Figures 2 and 4 and Table 2. The PCR reactions were performed using enhanced DNA polymerase, Elongase (Invitrogen, Carlsbad, CA) with ~1 min of extension

TABLE 2

Primers used to amplify maize *p1* sequences

Primer names	Primer sequences (5′–3′)
P1rr-18	TGAGTCCTGACCGACAGTCT
PA-B4	tgccttccatacttgcactgc
P1rr-14	GAAGGCAGACGATGAGGAGA
PA-B6	CACAACCTTTCACATACAGAG
EP5-16	cgagacttggctcctgt
P1rr-16	TCTCAGAGTATAGCAACAC
P1rr-13r	CTCATCAACGTGCTGTTCC
P1rr-30	CGTCGTCAAGAACTCAAGAT
P1rr-16r	GTGTTGCTATACTCTGAGA
P1rr-29	GGCTTGGTGCGTTGCTGA
723-5	TCTAGGCACTTTCTCGTG
723-3	GTAGAAATAAAGTCTGAGCA
P1rr-32	TGTAAACCGTGCTCACTG
P1rr-25	TGTAAACCGTGCCAGTGA

The orientation and approximate positions of these primers are shown in Figures 2 and 4.

time for every 1-kb fragment size. Optimal PCR parameters were followed as suggested by the manufacturer. To minimize the problems caused by PCR artifacts due to annealing of partial extension products to homologous regions in the template DNA pool, the PCR products from three independent reactions were mixed, purified using a gel extraction kit (QIAGEN, Valencia, CA), and sequenced directly from both directions. Sequencing reactions were provided by the Iowa State University DNA Sequencing Facility. The final sequences were inspected and assembled using the software package, Vector NTI Advance 9.0 (InforMax, Frederick, MD). 5′ and 3′ noncoding repeats of the newly sequenced *p1* alleles, *P1-rr1088*, *P1-rrCFS36*, *P1-rwCFS302*, and *P1-rwCFS342*, have GenBank accession nos. DQ160218–DQ160223.

Sequence analysis: Sequences from *P1-rr1088*, *P1-rrCFS36*, *P1-rwCFS302*, and *P1-rwCFS342* were aligned with those from *P1-rr4B2* and *P1-rw1077* using the software package, Vector NTI Advance 9.0 (InforMax). Phylogenetic analysis was carried out using a maximum-likelihood method in PAUP version 4.01 (Swofford 1998). Heuristic searching was conducted using a general time-reversible evolutionary model with estimated base frequencies and rate variation across sites modeled by gamma distribution. Support values for nodes on the maximum-likelihood tree were estimated with 500 bootstrap replicates (Felsenstein 1985). A 50% majority-rule consensus tree was generated using the sequence from teosinte as an outgroup. DNA diversity estimation and sliding-window analysis on the *p1* alleles were conducted using the program package, DnaSP 4.0 (Rozas and Rozas 1999). The number of substitutions per nucleotide site between the distinct *p1* alleles was estimated under the Kimura two-parameter (gamma) model as implemented in MEGA v. 2.1.

RESULTS

Characterization of *p1* alleles: More than 100 natural *p1* variations have been collected and classified according to their pigmentation patterns in kernel pericarp and cob glumes (Brink and Styles 1966). Previous studies on three distinct *p1* alleles indicated that the *p1* gene could contain either a single coding sequence, as in *P1-rr4B2* and *P1-rw1077* (Lechelt *et al.* 1989; Zhang and Peterson 2005), or a multiple-copy complex, as in *P1-wr[w22]* (Chopra *et al.* 1998). A DNA gel blot survey with *p1*-specific probes has been conducted on 24 *P1-rr*, 6 *P1-rw*, and 9 *P1-wr* alleles (Cocciolone *et al.* 2001; M. D. McMullen, personal communication). The results showed that all examined *P1-wr* alleles contain multiple copies of *p1* sequences, while all *P1-rw* alleles are single-copy genes; the *P1-rr* alleles could be either single copy or multiple copy. Because of the complicated nature of multiple-copy *p1* alleles (Chopra *et al.* 1998), we focused on the characterization of single-copy *p1* alleles in this study. According to the DNA gel blot patterns, the single-copy *p1* alleles, including 6 *P1-rw* and 9 *P1-rr* alleles, can be classified into six allelic types: three *P1-rr* types and three *P1-rw* types (Table 1; Figure 1, A and B). The *P1-rr4B2* and *P1-rw1077* allelic types have been characterized and compared previously (Lechelt *et al.* 1989; Sidorenko *et al.* 2000; Zhang and Peterson 2005). Representative alleles from the other four allelic types (*P1-rrCFS36*, *P1-rr1088*, *P1-rwCFS302*, and *P1-rwCFS342*, phenotypes are shown in Figure 1A) were chosen for further analysis.

Structural comparisons of *p1* alleles: Previous studies have shown that the coding regions of different *p1* alleles share high sequence similarity; and DNA polymorphisms in noncoding regulatory regions other than in coding regions could be responsible for phenotypic variations of the *p1* alleles (Lechelt *et al.* 1989; Sidorenko *et al.* 2000; Zhang and Peterson 2005). To examine sequence polymorphisms in noncoding regions of the single-copy *p1* alleles studied here, we used genomic PCR with *p1*-specific primers to amplify both 5′ and 3′ noncoding regions of the *P1-rr* and *P1-rw* alleles. In each allele, ~6 kb PCR products from both 5′ and 3′ regions were sequenced and assembled, and the sequence assemblies match DNA gel blot patterns (see materials and methods). The structures of the noncoding regions of all six single-copy *p1* alleles were compared as shown in Figure 2. Like those in *P1-rr4B2* and *P1-rw1077*, the 5′ and 3′ noncoding regions in the four newly sequenced *p1* alleles are also present as long direct repeats flanking the *p1* coding sequences. The repeat unit could extend as long as 6.3 kb (as in *P1-rw1077*) from a 1.1-kb region, termed fragment C, to a 15-bp small repeat near the transcription start site $(5'\text{-GCGGGAGTGCGGCCT-}3')_2$.

Structural and sequence comparisons between the simplex *p1* alleles revealed a number of DNA polymorphisms in both 5′ and 3′ noncoding direct repeats. In the 5′ repeats, the most notable polymorphic regions overlap with the previously identified enhancer regions (Figure 2). The distal enhancer region is located 5 kb upstream of the *p1* transcription start site and includes the 405-bp fragment 15 and a 666-bp sequence termed fragment 14 (Sidorenko *et al.* 2000; Zhang and Peterson 2005; Figure 2). Based on the copy number of fragment 15 and fragment 14 sequences in the enhancer-containing regions, three distinct structural types can be defined in six simplex *p1* alleles: (i) 1 copy of fragment 15 and 0 copy of fragment 14, *i.e.*, the enhancer-containing region contains only fragment 15, as shown in *P1-rwCFS302* and *P1-rwCFS342*; (ii) 1.5 copies of fragment 15 and 1 copy of fragment 14, *i.e.*, the enhancer-contaning region contains a partial copy of fragment 15 followed by fragment 14 and a full copy of fragment 15, as seen in *P1-rw1077*; (iii) 2 copies of fragment 15 and 1 copy of fragment 14, *i.e.*, the enhancer-containing region contains a full copy of fragment 15 followed by fragment 14 and an additional full copy of fragment 15, as shown in *P1-rr1088*, *P1-rrCFS36*, and *P1-rr4B2*. The type iii sequences can be further classified as two subtypes: (a) *P1-rrCFS36* and *P1-rr4B2*, in which a 1.6-kb transposon-like sequence (Sidorenko *et al.* 2000) is present in the midpoint of the upstream fragment 15; (b) *P1-rr1088*, in which the 1.6-kb sequence is absent (Figure 2). The 1.6-kb transposon is flanked by an 8-bp direct repeat (CCAGTGAG), which

may represent a target site duplication (TSD) generated upon insertion of the 1.6-kb transposon-like sequence. The *P1-rr1088* allele contains only a single copy of the 8-bp sequence and no evidence of a sequence alteration (footprint), which commonly accompanies transposon excision. We conclude that the polymorphism between the two subtypes reflects a sequence insertion, rather than a deletion.

In the 3′ noncoding regions, the sequences composing fragment 15 and fragment 14 are also highly variable. Similar to the 5′ noncoding sequences, the 3′ regions of the simplex *p1* alleles can also be classified as three distinct types: (i) 1 copy of fragment 14 and 1 copy of fragment 15 as shown in *P1-rrCFS36*, *P1-rr1088*, and *P1-rwCFS342*; (ii) 1.5 copies of fragment 15 and 1 copy of fragment 14 as seen in *P1-rw1077*; and (iii) 2 copies of fragment 14 and 2 copies of fragment 15 as shown in *P1-rr4B2* (equivalent to two copies of type i sequences arranged in direct orientation).

Sequence alignment also revealed two large deletions in the 5′ and 3′ noncoding regions of the *P1-rwCFS342* alleles: a 549-bp sequence (738 bp downstream of fragment 15) that is absent from both 5′ and 3′ noncoding direct repeats of *P1-rwCFS342* (Figure 2) and a 148-bp sequence (~3500 bp downstream of fragment 15 in the 3′ repeat, Figure 2) that is absent from the 3′ noncoding regions of *P1-rwCFS342*. Sequence analysis of a *p1* homologous gene from teosinte (*Zea mays* subsp. *parviglumis*) (Zhang *et al.* 2000), the immediate progenitor of modern maize, indicated that both 549- and 148-bp sequences are present in teosinte (data not shown). Possibly, these two polymorphic regions resulted from deletions in *P1-rwCFS342* rather than insertions in other *p1* alleles.

Variations in the distal enhancers correlate with changes of pigmentation patterns in cob glumes: Comparisons of the simplex *p1* alleles reveal three distinct structural types in the 5′ enhancer-containing region. In all the simplex *P1-rr* alleles, this region carries duplicated fragment 15 sequences as well as a fragment 14 sequence between them. A 386-bp cob glume-specific regulatory region has been identified from the *P1-rr4B2* allele, which overlaps with the upstream copy of fragment 15 plus the downstream 197-bp sequence

Figure 1.—(A) Phenotypes of the natural *p1* alleles. Mature ear pigmentation patterns specified by the *p1* alleles: *P1-rw1077*, *P1-rwCFS302*, and *P1-rwCFS342* (top row from left to right) and *P1-rr1088*, *P1-rrCFS36*, and *P1-rr4B2* (bottom row from left to right). All alleles are homozygous. (B) DNA gel blot analyses on the *p1* simplex alleles: lane 1, *P1-rw1077*; lane 2, *P1-rwCFS302*; lane 3, *P1-rwCFS342*; lane 4, *P1-rr4B2*; lane 5, *P1-rr1088*; and lane 6, *P1-rrCFS36*. Genomic DNA was digested with *Sal*I and hybridized with the probe, *p1* genomic fragment 15. The 1.2-kb band in *P1-rr4B2* is a doublet, while *P1-rr1088* and *P1-rrCFS36* have only one copy of the 1.2-kb fragment.

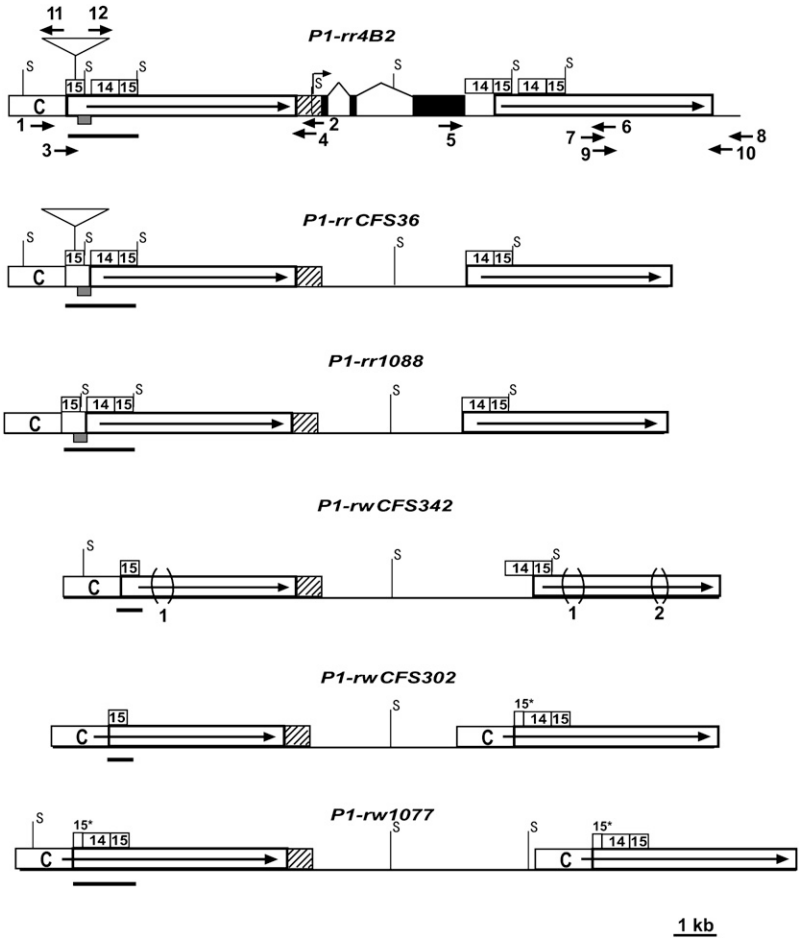

Figure 2.—Structural comparisons of the simplex *p1* alleles. The solid boxes connected by thin lines represent exon regions. The open boxes with solid arrows indicate long direct repeats in 5′ and 3′ noncoding regions. The bent arrow represents a transcription start site (Grotewold *et al.* 1991). The hatched boxes represent 5′-UTR and 90-bp promoter regions conserved among the *p1* alleles. The open boxes labeled with C indicate a 1.1-kb region, termed fragment C, present at the 5′ direct repeats of all *p1* alleles, but absent from the 3′ direct repeats in several *p1* alleles. The numbered open boxes represent the sequences identified in the text as fragment 15 and fragment 14. The enhancer-containing regions in each *p1* allele are highlighted by solid bars, which range from 1.6 kb (in the *P1-rr* alleles) to 602 bp (in the *P1-rwCFS302* and *P1-rwCFS342* alleles). The cob glume-specific regulatory region is indicated as shaded boxes in the enhancer-containing regions of the *P1-rr* alleles (which include the 189-bp sequence from fragment 15 and the 197-bp immediate downstream sequence). The regions marked 15* indicate partial fragment 15 sequences, which contain only the first 200 bp fragment 15 sequences. The triangles located in the first copy of fragment 15 in the 5′ noncoding repeats of *P1-rr4B2* and *P1-rrCFS36* represent insertions of a 1.6-kb transposon-like sequence. The bracketed regions marked 1 and 2 indicate 549- and 148-bp deletions, respectively, in the noncoding repeats of *P1-rwCFS342*. The solid arrows represent primers used to amplify 5′ and 3′ noncoding regions of *p1*: 1, P1rr-18; 2, PA-B4; 3, P1rr-14; 4, PA-B6; 5, EP5-16; 6, P1rr-16; 7, P1rr-13r; 8, P1rr-30; 9, P1rr-16r; 10, P1rr-29; 11, P1rr-32; and 12, P1rr-25. Positions of the restriction site, *Sal*I, are shown on each *p1* allele (only unmethylated *Sal*I sites that flank fragment 15 are indicated on the map).

(Figure 2). Although this particular sequence is present in both 5′ and 3′ direct noncoding repeats, the observation that the *P1-rw1077* allele lacks this region in that specific location at the enhancer-containing region suggests that this cob glume enhancer region functions in a position-specific manner (Zhang and Peterson 2005). Absence of the cob glume-specific enhancer in that specific location is also observed in the newly sequenced *P1-rw* alleles, *P1-rwCFS302* and *P1-rwCFS342*, which carry only one copy of fragment 15 (without duplication and fragment 14 sequence) in the enhancer-containing region (Figure 2). Therefore, in the single-copy *p1* alleles examined in this study, the presence/absence of the cob glume-specific regulatory region in a particular position is correlated with gain/loss of pigmentation in cob glumes. Moreover, structural comparisons between the *P1-rr* and *P1-rw* alleles indicated that the cob glume-specific region is formed by complete duplication of fragment 15 and insertion of fragment 14 sequences. Our further analysis suggested that gene conversion between 5′ and 3′ noncoding repeats could account for the duplication and insertion events observed in the enhancer-containing regions of the simplex *p1* alleles, which brought the cob glume-specific regulatory elements into the right position (see below).

Evidence for gene conversion between noncoding regions in the *p1* alleles: The 5′ and 3′ noncoding regions of the *p1* locus are arranged as direct repeats, separated by the ~6-kb transcribed region. The observation that a 549-bp deletion is located at identical positions in both 5′ and 3′ noncoding regions (*P1-rwCFS342*; Figure 2) suggests that gene conversion events occurred between the noncoding direct repeats in the *p1* locus. Gene conversion has been indicated as a candidate mechanism for both homogenizing and diversifying tandem repeats (Teshima and Innan 2004). As mentioned above, the enhancer-containing regions in the 5′ noncoding repeats have diverse structures in the distinct *p1* simplex alleles. The primary aim of this study is to understand how the 5′ enhancer-containing region was diversified and how the cob glume-specific sequence came to occupy its specific position in the *P1-rr* alleles. To assess the potential role of gene conversion in the creation of diversity in that region, phylogenetic analysis was performed on a 602-bp sequence, which includes fragment 15 and a 3′ adjacent 197-bp

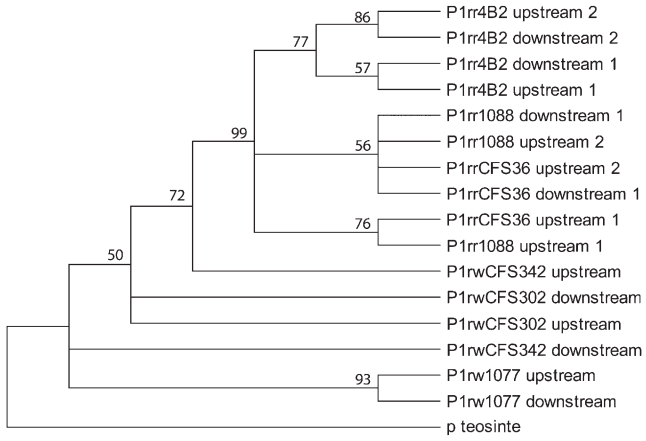

FIGURE 3.—Phylogenetic analysis of the *p1* distal enhancer regions. A rooted 50% majority-rule consensus maximum-likelihood tree was generated by using a 602-bp sequence, which includes fragment 15 plus a 197-bp downstream sequence. The sequence from the maize wild relative, *teosinte parviglumis*, was used as the outgroup. The numbers at nodes represent bootstrap values derived from 500 replicates. Sequences from the 5′ and 3′ noncoding repeats are identified as upstream or downstream, respectively. Sequences from the same repeat are indicated in numeric order from 5′ to 3′.

sequence. This 602-bp sequence is present in both 5′ and 3′ noncoding regions of all single-copy *p1* alleles and can be viewed as a small repeat unit. Thus, it will be more informative to use the entire 602-bp region in the phylogenetic analysis than to use only the 386-bp cob glume enhancer-containing region. In the *P1-rw* alleles, only one copy of the 602-bp sequence is present in the 5′ noncoding regions, while two *P1-rr* alleles (*P1-rr1088* and *P1-rrCSF36*) have duplicated 602-bp sequences in their 5′ repeats; and *P1-rr4B2* has the 602-bp sequence duplicated in both 5′ and 3′ repeats (Figure 2). A total of 17 sequences were used to generate a maximum-likelihood consensus tree (Figure 3): 16 sequences from 5′ and 3′ repeats of the simplex *p1* alleles and one sequence from the 3′ noncoding region of the teosinte *p* gene (*p2t*) used as the outgroup (ZHANG *et al.* 2000). The 1.6-kb transposon-like sequence and a flanking 8-bp TSD were removed from the 5′ 602-bp sequence of *P1-rr4B2* and *P1-rr1088* for this analysis. If the 5′ repeats evolved independently from the 3′ repeats, then the 5′ sequences and the 3′ sequences should form distinct groups in the phylogenetic tree. Gene conversion events between the 5′ and 3′ repeats, however, could change the phylogenetic patterns by clustering the 5′ sequence with the 3′ sequence from the same *p1* allele. On the basis of this logic, from the resulting phylogenetic tree, potential gene conversion events were detected between the 5′ and 3′ 602-bp sequences from four *p1* alleles, *i.e.*, *P1-rw1077*, *P1-rr4B2*, *P1-rrCFS36*, and *P1-rr1088*. Interestingly, the second copies (3′ copies) of the duplicated 602-bp sequences in the 5′ repeats of all *P1-rr* alleles are more closely related to the 602-bp sequences in the 3′ repeats than to their adjacent 5′ copies. This observation suggests that gene conversion events are involved in generating *P1-rr*-specific enhancer structures (see DISCUSSION). In contrast to the sequences from the *P1-rr* alleles, which form a monophyletic group with a high bootstrapping value (99%), those sequences from the *P1-rw* alleles are closely related to that from the *p* homologous gene in teosintes. This could suggest early divergence of the *P1-rw* alleles from the *p1* ancestral gene during evolution (see DISCUSSION).

Nucleotide diversity in *p1* noncoding regions: Nucleotide diversity among the simplex *p1* alleles was first estimated in both the 5′ and the 3′ noncoding repeats. As indicated in Figure 4, the regions that were sampled in this analysis are shared in all simplex *p1* alleles. In the 5′ noncoding regions, a total of 19 indels and 83 polymorphic sites were detected, while, in the 3′ noncoding regions, a total of 42 indels and 197 polymorphic sites were detected. The estimated values (π and θ) for DNA diversity also indicated that the 3′ regions are more diverse than the 5′ regions (*e.g.*, π-value 0.02929 *vs.* 0.00906; Table 3). This suggests that the 5′ noncoding repeats, but not the 3′ repeats, contain the critical regulatory regions for *p1* expression and are therefore under functional constraints (SIDORENKO *et al.* 2000).

To investigate the distribution of polymorphic sites, sliding-window analysis was performed on both 5′ and 3′ noncoding regions. This analysis revealed that the

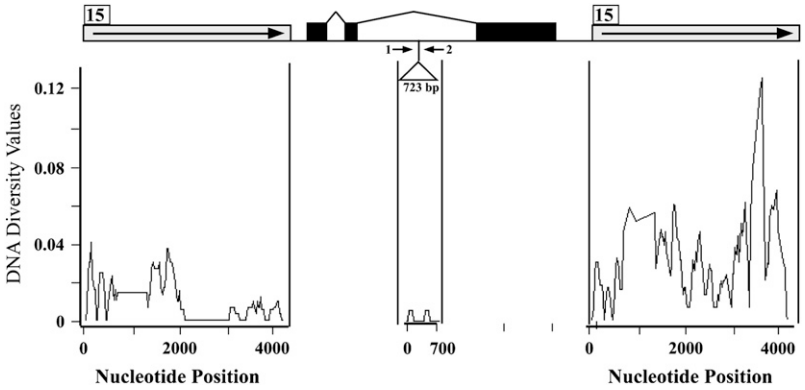

FIGURE 4.—Nucleotide diversity analysis of noncoding regions at the *p1* locus. Open boxes with solid arrows represent the 5′ and 3′ noncoding direct repeats. The position of fragment 15 is indicated with the numbered box. The solid boxes connected with thin lines represent the exon and intron structure of the *p1* locus. The triangle in the second intron indicates the 723-bp sequence that is also subject to nucleotide diversity analysis. The solid arrows represent the primers for amplification of the 723-bp region: 1, 723-5; 2, 723-3.

TABLE 3

Nucleotide diversity in the simplex *p1* alleles

Regions	*n*	No. of silent sites	No. of polymorphic sites	Θ	π	Tajima's *D*
Upstream repeat	6	3526	83	0.01031	0.00906	−0.7840, NS
Downstream repeat	6	3278	197	0.02483	0.02929	−1.1599, NS
723-bp sequence	5	723	2	0.00133	0.00111	−0.97256, NS

Nucleotide diversity was estimated on the simplex *p1* haplotypes. NS, not significant.

polymorphic sites are distributed throughout the 3′ noncoding regions, while, in the 5′ noncoding regions, polymorphic sites were frequently observed in the first 2-kbp region but are much less frequent in the remaining regions (Figure 4). Unequal functional constraints on the two regions could be one possible explanation for heterogeneity of nucleotide diversity in the 5′ noncoding regions. By *Ac* transposon mutagenesis assay, however, no functional elements were identified in the low-diversity region, while the distal enhancer was identified in the first 2-kb region (Moreno *et al.* 1992; Sidorenko *et al.* 2000). Alternatively, the uneven distribution of polymorphic sites in the 5′ noncoding regions could result from variation in recombination frequency across this region (see discussion).

Phylogenetic analysis has suggested that DNA diversity of 5′ and 3′ repeats at the simplex *p1* alleles could have been affected by gene conversion. To further test this idea, we compared nucleotide diversity between duplicated and nonduplicated *p1* sequences. An ideal single-copy *p1* sequence for this analysis is the 723-bp sequence located in the second intron of *p1* (Figure 4), because (i) this sequence is present only in the *p1* locus but not in the *p2* gene, the tightly linked paralogous gene of *p1*, which has potential to undergo gene conversion with *p1* (Zhang *et al.* 2000; Zhang and Peterson 2005), and (ii) this sequence seems to be subject to no functional constraints as it is dispensable for expression of *P1-rw1077* (Zhang and Peterson 2005). Interestingly, the values for DNA diversity of the 723-bp sequence (π-value is 0.00133 and θ-value is 0.00111) are 8–26 times lower than those of the 5′ and 3′ noncoding regions (Table 3). This result is consistent with the idea that homologous recombination, *e.g.*, gene conversion, between duplicated sequences increased the diversity in the *p1* noncoding repeats.

Estimated divergence time of the *p1* alleles: As stated above, the 723-bp sequence in the *p1* second intron likely evolved neutrally; and nucleotide diversity of this region could not have been affected by local gene conversion events. Thus, this sequence can be used to estimate the divergence time of the *p1* alleles. The pairwise nucleotide substitution rates were estimated between the simplex *p1* alleles (Table 4). Because *P1-rw1077* lacks the 723-bp sequence, the substitution rates between *P1-rw1077* and other *p1* alleles cannot be determined. On the basis of the published estimates of substitution rates in plant nuclear genes (ranging from 2.6×10^{-9} to 1.5×10^{-8} substitutions per synonymous site per year; Gaut 1998; Senchina *et al.* 2003), the divergence time of two *P1-rw* alleles, *P1-rwCFS342* and *P1-rwCFS302*, is estimated as 93,000–540,000 years ago, while the time between *P1-rw* (either *P1-rwCFS342* or *P1-rwCFS302*) and *P1-rr* is ~47,000–270,000 years ago (Figure 5). The time of divergence between the *P1-rr* alleles cannot be estimated, because no substitutions were detected between these alleles (using the formula $T = K/2r$, where K represents divergence amount and r equals the rate of mutation).

DISCUSSION

Origin and diversification of the *p1* gene: Alleles of the *p1* gene exhibit a great degree of genetic and

TABLE 4

Estimated number of substitutions per nucleotide site between the *p1* alleles

	P1-rw CFS302	*P1-rw CFS342*	*P1-rw 1077*	*P1-rr CFS36*	*P1-rr 1088*	P1-rr *4B2*
P1-rw CFS302	—					
P1-rw CFS342	0.0028 (2)	—				
P1-rw 1077	NA	NA	—			
P1-rr CFS36	0.0014 (1)	0.0014 (1)	NA	—		
P1-rr 1088	0.0014 (1)	0.0014 (1)	NA	0 (0)	—	
P1-rr 4B2	0.0014 (1)	0.0014 (1)	NA	0 (0)	0 (0)	—

Numbers in parentheses are the numbers of nucleotide substitutions over 723 sites.

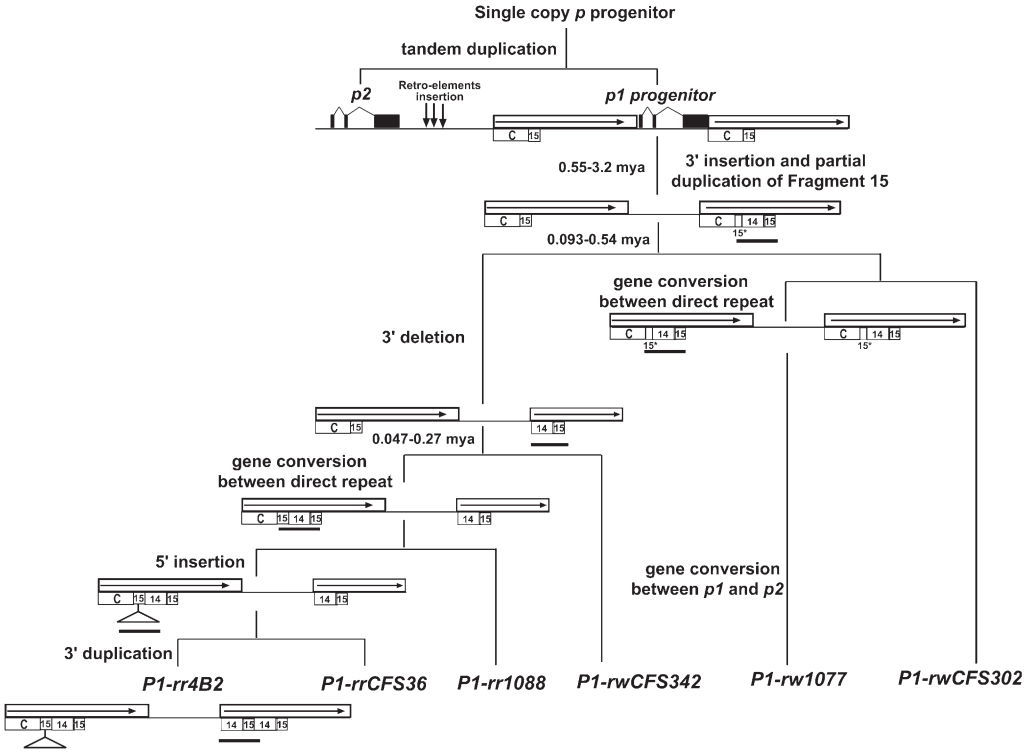

FIGURE 5.—Sequential model for *p1* evolution. The stepwise progression from *p* progenitor gene to the distinct *p1* alleles is shown. The open boxes with solid arrows indicated the noncoding direct repeats flanking the *p1* coding sequence. The numbered open boxes are the same as those shown in Figure 2. The solid boxes indicate exons of the *p1* and *p2* genes. The vertical arrows between the *p1* and *p2* regions represent retroelement insertions that separated the *p1* gene from *p2* (ZHANG *et al.* 2000). In each step, the rearranged regions are highlighted by solid bars. The divergence time between the *p1* alleles is indicated at the branch points.

phenotypic diversity (COCCIOLONE *et al.* 2001; ZHANG and PETERSON 2005). In a previous study, ZHANG *et al.* (2000) suggested that the *p1* gene was generated by a recent tandem duplication event. How was the observed high degree of diversity created at the *p1* locus in this short period of evolutionary time? In this study, investigation of six distinct *p1* alleles suggested that the noncoding repeats flanking the *p1* coding sequence may have played an important role in diversification of the *p1* locus during evolution.

Sequence comparisons of six distinct single-copy *p1* alleles have revealed that they all contain the long direct noncoding repeats that flank the *p1* coding sequence. The gene structure of the common ancestor of all *p1* alleles has been deduced on the basis of analysis of a *p*-homologous gene in teosinte (ZHANG *et al.* 2000). As indicated in Figure 5, in the *p1* progenitor gene, the long direct repeat is ~6 kb and extends from a 5′ sequence termed fragment C to a small 15-bp repeat. Subsequent sequence rearrangements in both 5′ and 3′ repeats diversified the *p1* alleles (Table 3).

Structural comparisons of the distinct *p1* alleles as well as evidence from phylogenetic analysis and nucleotide diversity analysis allowed us to propose a stepwise evolutionary model to most parsimoniously account for generation of the distinct single-copy *p1* alleles. In this model, as indicated in Figure 5, the DNA rearrangements started at the 3′ noncoding repeats with insertion of fragment 14 and partial duplication of fragment 15, *i.e.*, the structure shown in *P1-rwCFS302*. Such rearrangements could be transferred to the 5′ noncoding region by gene conversion as shown in *P1-rw1077*; or they can be further modified by deletion, giving rise to the structures of the 3′ noncoding regions observed in the *P1-rwCFS342* and *P1-rr* alleles (Figures 2 and 5).

Because of low nucleotide polymorphisms in the enhancer-containing region, the relationship between the simplex *p1* alleles is not well resolved in the maximum-likelihood tree. Although our proposed evolution model is not in agreement with the phylogenetic tree in every detail, both scenarios state that the *P1-rw* alleles were present early in the evolutionary pathway, while the *P1-rr* alleles likely originated more recently. The divergence time of the *p1* alleles, estimated on the basis of the 723-bp sequence from the *p1* second intron (Figure 4), seems to postdate the estimated birth date of the *p1* progenitor gene (2.75 MYA; ZHANG *et al.* 2000). This is consistent with the hypothesis that all *p1* alleles studied to date are derived from tandem duplication of a single *p1* ancestral gene (ZHANG *et al.* 2000). On the other hand, the diversification of the *p1* gene appears to predate domestication of maize from teosinte ~7500 years ago (ILTIS 1983). Most likely, introgression of various *p1* alleles from teosinte into the domesticated maize gene pool was facilitated by positive human selection for the obvious pigmentation phenotypes. However, our estimation does not preclude the possibility that divergence of the *p1* alleles occurred during or after maize domestication, as the DNA diversity of the *p1* alleles is very low. Further investigation of *p1*-homologous genes in representative teosinte stocks is needed to more precisely estimate the divergence time.

Gene conversion could promote diversification of the *p1* noncoding repeats: In our model, the most interesting step in the evolution of the simplex *p1* locus is the change from the *P1-rwCFS342*-like structure, which contains only one copy of fragment 15 in the enhancer-containing region, to a *P1-rr*-like structure that has two copies of fragment 15 and one copy of fragment 14 in that region (Figures 2 and 5). The phylogenetic analysis based on the sequences containing the distal enhancer indicated that the second copy of the duplicated fragment 15 sequences at the 5′ noncoding region is more closely related to fragment 15 in the 3′ noncoding region rather than to the 5′ duplicated copy. These results suggest that the 5′ *P1-rr*-type structure in the enhancer-containing region could be generated by gene conversion between the 5′ and 3′ repeats. More specifically, we propose that, in an allele with a structure similar to that of *P1-rwCFS342*, fragment 14 and 15 sequences were converted from the 3′ repeat into the 5′ repeat, with the original fragment 15 sequence adjoined to the newly transferred sequence. After gene conversion, the enhancer-containing region of the resultant *p1* allele acquired an extra fragment 15 sequence as well as a fragment 14 sequence downstream of the original fragment 15. It has been suggested that gene conversion can create a new mosaic sequence by transferring a segment of differential sequence from a homologous donor region (Martinsohn *et al.* 1999). In this study, the conversion-generated mosaic sequences in the *P1-rr* alleles contain duplicated fragment 15 sequences with insertion of a fragment 14 sequence. As the result, a novel cob glume-specific regulatory region was created in the distal enhancer of *P1-rr*.

Diversifying gene conversion events between duplicated sequences could also explain the increased DNA diversity in the *p1* noncoding repeats compared to the single-copy sequence within the second intron. Recent studies have identified gene conversion as a major force to generate nucleotide diversity between homologous sequences (Ohta 1995; Angers *et al.* 2002; Nielsen *et al.* 2003; Teshima and Innan 2004; Backstrom *et al.* 2005). It has been suggested that gene conversion could be an error-prone process due to biased tendency toward GC in repair systems as well as a higher rate of misincorporation relative to replicative DNA synthesis (Martinsohn *et al.* 1999; Marais 2003). In addition, polarity gradients of gene conversion rates have been reported for conversion tracts in yeast (Martinsohn *et al.* 1999). Unequal gene conversion rates may explain the uneven distribution of polymorphisms in the 5′ *p1* noncoding repeats. For example, gene conversion rates may be higher in the first 2-kbp region than in the further downstream regions. As a result, DNA diversity in the first 2-kbp region in the 5′ noncoding region could have increased by converting nucleotide variations present in the 3′ noncoding regions, which probably accumulated more freely there due to fewer functional constraints. Uneven distribution of recombination events has also been reported in *a1*, *b1*, and *r1* loci in maize (Eggleston *et al.* 1995; Patterson *et al.* 1995; Yao *et al.* 2002). Because of the small sample size (six *p1* alleles) used in this study, however, investigation of additional *p1* variations is necessary to test these hypotheses.

Evolution of the distal enhancer and phenotypic variations at the *p1* locus: The stepwise evolution model proposed that the *p1* distal enhancer evolved from a simple to complex structures, *i.e.*, from a single copy of fragment 15 to duplicated fragment 15 sequences with insertion of fragment 14. In a recent study (Zhang and Peterson 2005), a cob glume-specific regulatory region was identified within the complex enhancer types (Figure 2). This particular structure is associated with all *P1-rr* alleles, but is absent from the *P1-rw* alleles. Thus, the evolution of the distal enhancer region appears to accompany the evolution of phenotypic variations observed at the *p1* gene. We propose that the early *p1* alleles conferred the red kernel pericarp and white cob glumes (RW) phenotype and that a later event may have converted the distal enhancer into a type that contains the cob glume-specific regulatory region. Such change resulted in acquisition of *p1* expression in cob glumes and thereby gave rise to the red kernel pericarp and red cob glumes (RR) phenotype.

Taken together, these results suggest that gene conversion between noncoding repeat sequences could be a major driving force for generation of genetic and phenotypic diversity at the *p1* locus. Given that tandem and segmental duplications are common in both animal and plant genomes (Blanc *et al.* 2000; McLysaght *et al.* 2002; Yu *et al.* 2005), recombination between noncoding duplicated sequences could have a major impact on molecular evolution and diversity in gene expression.

We thank Johnathan Wendel, Xun Gu, Erik Vollbrecht, Diane Bassham, Steve Whitham, and Dan Voytas for advice and comments on this manuscript. This material is based upon work supported by the National Science Foundation under grant no. 9601285. This journal article of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, was supported by Hatch Act and State of Iowa funds.

LITERATURE CITED

Angers, B., K. Gharbi and A. Estoup, 2002 Evidence of gene conversion events between paralogous sequences produced by tetraploidization in Salmoninae fish. J. Mol. Evol. **54:** 501–510.

Backstrom, N., H. Ceplitis, S. Berlin and H. Ellegren, 2005 Gene conversion drives the evolution of HINTW, an ampliconic gene on the female-specific avian W chromosome. Mol. Biol. Evol. **22:** 1992–1999.

Blanc, G., A. Barakat, R. Guyot, R. Cooke and M. Delseny, 2000 Extensive duplication and reshuffling in the Arabidopsis genome. Plant Cell **12:** 1093–1102.

Brink, R., and D. Styles, 1966 A collection of pericarp factors. Maize Genet. Coop. Newsl. **40:** 149–160.

Carroll, S. B., J. K. Grenier and S. D. Weatherbee, 2004 *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design.* Blackwell Science, Malden, MA.

Chopra, S., P. Athma, X. G. Li and T. Peterson, 1998 A maize Myb homolog is encoded by a multicopy gene complex. Mol. Gen. Genet. **260:** 372–380.

Cocciolone, S. M., S. Chopra, S. A. Flint-Garcia, M. D. McMullen and T. Peterson, 2001 Tissue-specific patterns of a maize Myb transcription factor are epigenetically regulated. Plant J. **27:** 467–478.

Cone, K. C., S. M. Cocciolone, F. A. Burr and B. Burr, 1993 Maize anthocyanin regulatory gene pl is a duplicate of c1 that functions in the plant. Plant Cell **5:** 1795–1805.

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

Eggleston, W. B., M. Alleman and J. L. Kermicle, 1995 Molecular organization and germinal instability of R-stippled maize. Genetics **141:** 347–360.

Felsenstein, J., 1985 Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39:** 783–791.

Gaut, B. S., 1998 Molecular clocks and nucleotide substitution rates in higher plants, pp. 93–120 in *Evolutionary Biology*, edited by M. K. Hecht, W. C. Steere and B. Wallace. Plenum, New York.

Grotewold, E., P. Athma and T. Peterson, 1991 Alternatively spliced products of the maize *P* gene encode proteins with homology to the DNA binding domain of Myb-like transcription factors. Proc. Natl. Acad. Sci. USA **88:** 4587–4591.

Hanson, M. A., B. S. Gaut, A. O. Stec, S. I. Fuerstenberg, M. M. Goodman *et al.*, 1996 Evolution of anthocyanin biosynthesis in maize kernels: the role of regulatory and enzymatic loci. Genetics **143:** 1395–1407.

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

Jiang, C., J. Gu, X. Gu, S. Chopra and T. Peterson, 2004 Ordered origin of the typical two- and three-repeat Myb genes. Gene **326:** 13–22.

Lechelt, C., T. Peterson, A. Laird, J. Chen, S. L. Dellaporta *et al.*, 1989 Isolation and molecular analysis of the maize P locus. Mol. Gen. Genet. **219:** 225–234.

Li, Y., J. P. Bernot, C. Illingworth, W. Lison, K. M. Bernot *et al.*, 2001 Gene conversion within regulatory sequences generates maize r alleles with altered gene expression. Genetics **159:** 1727–1740.

Ludwig, M. Z., 2002 Functional evolution of noncoding DNA. Curr. Opin. Genet. Dev. **12:** 634–639.

Ludwig, M. Z., C. Bergman, N. H. Patel and M. Kreitman, 2000 Evidence for stabilizing selection in a eukaryotic enhancer element. Nature **403:** 564–567.

Ludwig, S. R., and S. R. Wessler, 1990 Maize R gene family: tissue-specific helix-loop-helix proteins. Cell **62:** 849–851.

Marais, G., 2003 Biased gene conversion: implications for genome and sex evolution. Trends Genet. **19:** 330–338.

Martinsohn, J. T., A. B. Sousa, L. A. Guethlein and J. C. Howard, 1999 The gene conversion hypothesis of MHC evolution: a review. Immunogenetics **50:** 168–200.

McLysaght, A., K. Hokamp and K. H. Wolfe, 2002 Extensive genomic duplication during early chordate evolution. Nat. Genet. **31:** 200–204.

Moreno, M. A., J. Chen, I. Greenblatt and S. L. Dellaporta, 1992 Reconstitutional mutagenesis of the maize P gene by short-range Ac transpositions. Genetics **131:** 939–956.

Nielsen, K. M., J. Kasper, M. Choi, T. Bedford, K. Kristiansen *et al.*, 2003 Gene conversion as a source of nucleotide diversity in Plasmodium falciparum. Mol. Biol. Evol. **20:** 726–734.

Ohta, T., 1995 Gene conversion vs point mutation in generating variability at the antigen recognition site of major histocompatibility complex loci. J. Mol. Evol. **41:** 115–119.

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.

Procissi, A., S. Dolfini, A. Ronchi and C. Tonelli, 1997 Light-dependent spatial and temporal expression of pigment regulatory genes in developing maize seeds. Plant Cell **9:** 1547–1557.

Purugganan, M. D., 2000 The molecular population genetics of regulatory genes. Mol. Ecol. **9:** 1451–1461.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

Saghai-Maroof, M. A., K. M. Soliman, R. A. Jorgensen and R. W. Allard, 1984 Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proc. Natl. Acad. Sci. USA **81:** 8014–8018.

Selinger, D. A., D. Lisch and V. L. Chandler, 1998 The maize regulatory gene B-Peru contains a DNA rearrangement that specifies tissue-specific expression through both positive and negative promoter elements. Genetics **149:** 1125–1138.

Senchina, D. S., I. Alvarez, R. C. Cronn, B. Liu, J. Rong *et al.*, 2003 Rate variation among nuclear genes and the age of polyploidy in Gossypium. Mol. Biol. Evol. **20:** 633–643.

Sidorenko, L. V., X. Li, S. M. Cocciolone, S. Chopra, L. Tagliani *et al.*, 2000 Complex structure of a maize Myb gene promoter: functional analysis in transgenic plants. Plant J. **22:** 471–482.

Swofford, D. L., 1998 *PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods)*. Sinauer Associates, Sunderland, MA.

Teshima, K. M., and H. Innan, 2004 The effect of gene conversion on the divergence between duplicated genes. Genetics **166:** 1553–1560.

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

Wray, G. A., M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer *et al.*, 2003 The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. **20:** 1377–1419.

Yao, H., Q. Zhou, J. Li, H. Smith, M. Yandeau *et al.*, 2002 From the cover: molecular characterization of meiotic recombination across the 140-kb multigenic a1-sh2 interval of maize. Proc. Natl. Acad. Sci. USA **99:** 6157–6162.

Yu, J., J. Wang, W. Lin, S. Li, H. Li *et al.*, 2005 The genomes of Oryza sativa: a history of duplications. PLoS Biol. **3:** e38.

Zhang, F., and T. Peterson, 2005 Comparisons of maize pericarp color1 alleles reveal paralogous gene recombination and an organ-specific enhancer region. Plant Cell **17:** 903–914.

Zhang, P., S. Chopra and T. Peterson, 2000 A segmental gene duplication generated differentially expressed myb-homologous genes in maize. Plant Cell **12:** 2311–2322.

Communicating editor: J. A. Birchler