

## Excision of an Active CACTA-Like Transposable Element From *DFR2* Causes Variegated Flowers in Soybean [*Glycine max* (L.) Merr.]

Min Xu,\* Hargeet K. Brar,\* Sehiza Grosic,\* Reid G. Palmer<sup>†</sup> and Madan K. Bhattacharyya<sup>\*,\*1</sup>

\*Department of Agronomy, Iowa State University, Ames, Iowa 50011 and <sup>†</sup>United States Department of Agriculture, Agricultural Research Service, Corn Insects and Crop Genetics Research, Department of Agronomy, Iowa State University, Ames, Iowa 50011

Manuscript received July 29, 2009  
Accepted for publication November 1, 2009

### ABSTRACT

Active endogenous transposable elements, useful tools for gene isolation, have not been reported from any legume species. An active transposable element was suggested to reside in the *W4* locus that governs flower color in soybean. Through biochemical and molecular analyses of several revertants of the *w4-m* allele, we have shown that the *W4* locus encodes dihydroflavonol-4-reductase 2 (*DFR2*). *w4-m* has arisen through insertion of *Tgm9*, a 20,548-bp CACTA-like transposable element, into the second intron of *DFR2*. *Tgm9* showed high nucleic acid sequence identity to *Tgmt\**. Its 5' and 3' terminal inverted repeats start with conserved CACTA sequence. The 3' subterminal region is highly repetitive. *Tgm9* carries *TNP1*- and *TNP2*-like transposase genes that are expressed in the mutable line, T322 (*w4-m*). The element excises at a high frequency from both somatic and germinal tissues. Following excision, reinsertions of *Tgm9* into the *DFR2* promoter generated novel stable alleles, *w4-dp* (dilute purple flowers) and *w4-p* (pale flowers). We hypothesize that the element is fractured during transposition, and truncated versions of the element in new insertion sites cause stable mutations. The highly active endogenous transposon, *Tgm9*, should facilitate genomics studies specifically that relate to legume biology.

**I**N soybean [*Glycine max* (L.) Merr.], five loci *W1*, *W3*, *W4*, *Wm*, and *Wp* control the pigmentations in flowers and hypocotyls (PALMER *et al.* 2004). Soybean plants with genotype *W1\_ w3w3 W4\_ Wm\_ Wp\_* produce wild-type purple flowers (Figure 1) and purple hypocotyls. Mutations at the *W4* locus in the *W1\_* background result in altered pigment accumulation patterns in petals and reduced levels of purple pigments in flowers and hypocotyls. Four mutant alleles, *w4*, *w4-m*, *w4-dp*, and *w4-p* have been mapped to this locus. The *w4* allele represents a spontaneous mutation, which produces near-white flowers (Figure 1) and green hypocotyls (HARTWIG and HINSON 1962; GROOSE and PALMER 1991). The *w4-m* allele was identified from a cross between two experimental breeding lines with white and purple flowers, respectively (PALMER *et al.* 1989; WEIGELT *et al.* 1990). *w4-m* is characterized by variegated flowers (Figure 1) and green hypocotyls with purple sectors (GROOSE *et al.* 1988).

*w4-m* has been proposed to harbor a class II transposable element (PALMER *et al.* 1989). Presumably, somatic

excision of the putative transposable element results in the variegated (GROOSE *et al.* 1988) and germinal excision wild-type phenotypes, purple flowers and purple pigments on hypocotyls (PALMER *et al.* 1989; GROOSE *et al.* 1990). The mutable line carrying *w4-m* undergoes germinal reversion at a very high frequency, about 6% per generation (GROOSE *et al.* 1990). Approximately 1% of the progeny derived from germinal revertants contain new mutations in unlinked loci, presumably resulting from reinsertion of the element (PALMER *et al.* 1989). For example, female partial-sterile 1 (*Fsp1*), female partial-sterile 2 (*Fsp2*), female partial-sterile 3 (*Fsp3*), and female partial-sterile 4 (*Fsp4*) were isolated from progenies of germinal revertants with purple flowers and were mapped to molecular linkage groups (MLG) C2, A2, F, and G, respectively (KATO and PALMER 2004). Similarly, 36 male-sterile, female-sterile mutants mapped to the *st8* region on MLG J (KATO and PALMER 2003; PALMER *et al.* 2008a), 24 necrotic root (*rn*) mutants mapped to the *rn* locus on MLG G (PALMER *et al.* 2008b), and three *Mdh1-n y20* mutants, mapped to a chromosomal region on MLG H (PALMER *et al.* 1989; XU and PALMER 2005b), were isolated among progenies of germinal revertants.

In addition to germinal revertants with purple flowers, the *w4* mutable line also generated intermediate stable revertants that produce flowers with variable pigment intensities ranging from purple to near-white (Figure 1). Two stable intermediate revertants, *w4-dp*

Supporting information is available online at <http://www.genetics.org/cgi/content/full/genetics.109.107904/DC1>.

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. DQ026299, EF187612, EU068464, and GQ344503.

<sup>1</sup>Corresponding author: Department of Agronomy, G303 Agronomy Hall, Iowa State University, Ames, IA 50011.  
E-mail: mbhattac@iastate.edu






Harosoy	Harosoy <i>w4</i>	T369	T321	T322
<i>W4W4</i>	<i>w4w4</i>	<i>w4-pw4-p</i>	<i>w4-dpw4-dp</i>	<i>w4-mw4-m</i>
Purple	Near white	Pale	Dilute purple	Variegated
				

FIGURE 1.—Variation in flower color among soybean lines carrying different *W4* alleles.

and *w4-p*, are allelic to *W4*. Plants carrying *w4-dp* or *w4-p* alleles produce dilute purple flowers or pale flowers, respectively (Figure 1) (PALMER and GROOSE 1993; XU and PALMER 2005a).

Pigment formation requires two types of genes: structural genes that encode anthocyanin biosynthetic enzymes [*e.g.*, CHS (chalcone synthase), F3H (flavanone 3-hydroxylase), DFR (dihydroflavonol-4-reductase), ANS (anthocyanidin synthase); Figure S1] and regulatory genes that control expression of structural genes (HOLTON and CORNISH 1995). Among the five genes, *W1*, *W3*, *W4*, *Wp*, and *Wm*, controlling pigment biosynthesis in soybean, four have been characterized at the molecular level (Figure S1). *W1* encodes a flavonoid 5', 3'-hydroxylase (ZABALA and VODKIN 2007). *W3* cosegregates with a *DFR* gene, *Wp* encodes a flavanone 3-hydroxylase (F3H), and *Wm* encodes a flavonol synthase (FLS) (FASOULA *et al.* 1995; ZABALA and VODKIN 2005; TAKAHASHI *et al.* 2007).

Nine CACTA-type class II transposable elements, *Tgm1*, *Tgm2*, *Tgm3*, *Tgm4*, *Tgm5*, *Tgm6*, *Tgm7*, *Tgm-Express1*, and *Tgm<sup>tr</sup>*, have been reported in soybean (RHODES and VODKIN 1988; ZABALA and VODKIN 2005, 2008). *Tgm-Express1* causes mutation in *Wp* (ZABALA and VODKIN 2005) and *Tgm<sup>tr</sup>* (EU190440) in *T* that encodes a flavonoid 3' hydroxylase (F3'H) (ZABALA and VODKIN 2003, 2008). The objectives of the present study were to characterize the *W4* locus and then investigate whether the *w4-m* allele harbors an active transposable element. Our results showed that a CACTA-like transposable element located in a dihydroflavonol-4-reductase gene causes variegated flower phenotype in soybean.

## MATERIALS AND METHODS

**Primers and probes:** All the primers and probes used in this study are listed in supporting information, Table S1 and Table S2, respectively.

**Plant materials:** Soybean lines differing for *W4* alleles were planted at the Bruner Farm, the United States Department of Agriculture (USDA) greenhouse or growth cabinet, Iowa State University (Ames, IA). Their genotypes and phenotypes are described in Table 1. For analyses of anthocyanins, flavonols, and RNAs, petals were collected from floral buds 1 day before

anthesis. For DNA analyses, genomic DNA was extracted from young leaves.

**Extraction and analysis of anthocyanins:** To extract anthocyanin pigments, freeze-dried flower petals were incubated in 1% (v/v) HCl in methanol for 3 hr at room temperature and centrifuged at 13,000 rpm for 10 min. Half of the supernatants was used for spectrophotometric analysis in a Beckman DU 640 nucleic acid and protein analyzer. The other half was hydrolyzed by boiling for 30 min. Hydrolyzed extracts were subjected to spectrophotometric analyses. The anthocyanidin contents were expressed as the absorbance at 535 nm ( $A_{535}$ ) per milligram of dried petals per milliliter of solvent.

**High performance liquid chromatography (HPLC) analysis of flavonols:** The flavonol aglycone samples of soybean flowers and authentic standard solutions of myricetin, quercetin, and kaempferol (Sigma, St. Louis, MO) were prepared according to BURBULIS *et al.* (1996), and stored at  $-20^{\circ}$ . Samples (100  $\mu$ l) were injected into a C-18 RP column attached to a Waters gradient HPLC system (Millipore, Billerica, MA) and eluted at a flow rate of 1.0 ml/min using the following linear gradient of HPLC-grade acetonitrile in HPLC-grade H<sub>2</sub>O (pH 3.0, adjusted with glacial acetic acid): 0 to 0% for 5 min, 0 to 10% for 5 min, 10 to 30% for 60 min, 30 to 100% for 5 min, 100 to 100% for 2 min, 100 to 0%, for 2 min, and 0 to 0% for 5 min. The system was run and data were acquired using Waters Millennium software, version 3.2. Elutents were analyzed by a photodiode array 996 detector (PDA996) at 255 nm and quantified by comparing to authentic standards.

**RNA preparation, RT-PCR, and RNA blot analysis:** Total RNA was prepared from immature petals using RNeasy mini kit (QIAGEN, Valencia, CA). cDNAs were synthesized from 2  $\mu$ g total RNAs using oligo-dT and SuperScript II reverse transcriptase (Invitrogen, Carlsbad, CA) and diluted twofold for PCR. Primers for PCR are listed in Table S1. For RNA blot analyses, 20  $\mu$ g total RNAs was separated on a 1.0% formaldehyde-agarose gel and blotted onto a Zeta Probe Nylon membrane (BioRad, Hercules, CA) by capillary transfer.

**DNA preparation and DNA blot analysis:** Genomic DNA was extracted from young leaves by following the CTAB method (KEIM *et al.* 1988), purified with equal volumes of phenol, phenol/chloroform (1:1/v/v), and chloroform (SAMBROOK *et al.* 1989). For DNA blot analysis, 10  $\mu$ g genomic DNA was digested with desired restriction enzymes and separated on a 0.8% (w/v) agarose gel. DNA blot analysis was conducted as previously described (SAMBROOK *et al.* 1989).

**BAC library screening and *W4* gene cloning:** A BAC library (BHATTACHARYYA *et al.* 2005) was screened using a partial *DFR* cDNA probe (Table S2). Positive clones were confirmed by DNA blot analysis. Sequence of the full-length *DFR2* gene was obtained through primer walking sequencing method. The BAC DNA for sequencing was extracted using the QIAGEN large constructs miniprep kit.

TABLE 1  
Soybean lines used in this study

Soybean lines	Genotypes	Flower color	Description of the <i>W4</i> locus
Harosoy	<i>W1W1 w3w3 WmWm WpWp W4W4</i>	Purple	Wild-type
T321	<i>W1W1 w3w3 WmWm WpWp w4-dpw4-dp</i>	Dilute purple	Mutant revertant from T322 ( <i>w4-m</i> )
T322	<i>W1W1 w3w3 WmWm WpWp w4-mw4-m</i>	Variegated	Mutable allele
T369	<i>W1W1 w3w3 WmWm WpWp w4-pw4-p</i>	Pale	Mutant revertant from T322 ( <i>w4-m</i> )
T325	<i>W1W1 w3w3 WmWm WpWp W4W4</i>	Purple	Wild-type revertant from T322 ( <i>w4-m</i> )
Williams	<i>w1w1 w3w3 WmWm WpWp W4W4</i>	White	Wild-type
Minsoy	<i>W1W1 w3w3 WmWm WpWp W4W4</i>	Purple	Wild-type
Williams 82	<i>w1w1 w3w3 WmWm WpWp W4W4</i>	White	Wild-type

**Genomic library construction and screening:** Two genomic libraries were constructed in the Lambda FIXII/*Xho*I vector (Stratagene, La Jolla, CA) using the DNA prepared from leaves of the T322 line homozygous for *w4-m* (PALMER *et al.* 1990). The DNA from the libraries was transferred to 137-mm nitrocellulose disks (Stratagene) (SAMBROOK *et al.* 1989). Approximately 0.4 million plaques of the first library and 1.5 million plaques of the second library were screened with a *DFR2* cDNA fragment (Table S2). Positive clones were confirmed by Southern blot analysis, PCR, and sequencing. The lambda DNA for sequencing was extracted using the QIAGEN Lambda Midi kit.

**PCR conditions:** PCR reactions were conducted in a 25- $\mu$ l mixture containing ~100 ng of genomic DNA or ~1 ng of plasmid or phage DNA or 2  $\mu$ l of first strand cDNA, 1 $\times$  PCR buffer, 2.0 mM MgCl<sub>2</sub>, 100  $\mu$ M dNTP, 0.15  $\mu$ M of each primer, and 1 unit of *Biolase Taq* polymerase (Bioline USA, Randolph, MA). PCR was started with an initial 2-min denaturation step at 94° followed by 5 cycles of 94° (30 sec), 60° (1 min, reduced by -1°/cycle), and 72° (1.5 min), and then by 27 cycles of 94° (30 sec), 54° (30 sec), and 72° (30 sec), with a final extension at 72° for 10 min.

**DNA sequencing and sequence analysis:** All the sequencing projects were conducted in an ABI 3730 DNA analyzer at the Iowa State University DNA facility. The local alignments were performed using BLAST (bl2seq) from NCBI (<http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi>). The global alignments and multiple alignments were conducted using ClustalW2 from EBI ([www.ebi.ac.uk/clustalw2](http://www.ebi.ac.uk/clustalw2)). Gene prediction was performed with GENSCAN (<http://genes.mit.edu/GENSCAN.html>). Polypeptide sequences were deduced from the DNA sequence using ExPASy translate tool (<http://ca.expasy.org/tools/dna.html>). Conserved domains in protein were searched with CDS program of NCBI (<http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>).

**Accession numbers:** Sequence data can be found in GenBank/EMBL database with accession nos. DQ026299 (*DFR2* partial cDNA), EF187612 (*DFR2* genomic sequence), EU068464 (truncated *Tgm9*), EU068463 (insertion in *w4-dp*), and GQ344503 (*Tgm9*).

## RESULTS

**The *w4* mutation blocks conversion of dihydromyricetin to delphinidin-3-monoglucoside:** The anthocyanins and flavonols present in flowers of four soybean lines, Harosoy (*W4*), T322 (*w4-m*), T321 (*w4-dp*), and T369 (*w4-p*), were investigated (Table 1). Anthocyanin extracts showed the maximum absorption peak at

535 nm with  $\lambda_{\max}$  450–650 nm (Figure S2). The peak shifted to 543 nm when the extracts were hydrolyzed by boiling (Figure S2). These spectral characteristics suggested that the main pigment in soybean flowers could be delphinidin-3-monoglucoside or its derivatives, petunidin-3-monoglucoside, and malvidin-3-monoglucoside (HARBORNE 1958), similar to the main pigment malvidin in soybean hypocotyls, stem, and subepidermal tissues (NOZZOLILLO 1973). Malvidin is generated through glycosylation and methylation of delphinidin. The anthocyanin contents in flower petal samples were investigated at 535 nm. The highest anthocyanin level was observed in wild-type purple petals (Harosoy) and purple petal sectors of T322, followed by pale flowers (T369) and dilute purple flowers (T321). The lowest anthocyanin content was observed in white petal sectors of T322 (Figure 2A).

Delphinidin-3-monoglucoside or its derivatives are believed to be the main pigments in soybean flowers. The flavonol myricetin is synthesized from a precursor of delphinidin-3-monoglucoside, dihydromyricetin by the enzyme flavonol synthase (FLS; Figure S1). HPLC analyses revealed enhanced accumulation of myricetin in petals of T321 and T369, and white petal sectors of T322 (Figure 2B) that showed less anthocyanin pigment accumulation (Figure 2B). These results suggested that the lesion in *w4* mutants is from dihydromyricetin to delphinidin-3-monoglucoside (Figure S1).

**Mutations in the *W4* locus were associated with reduced *DFR2* transcript levels:** We analyzed the *w4* mutants for steady state transcript levels of three structural genes, *F3H*, *DFR*, and *ANS* (Figure S1). The probes for *F3H* and *ANS* were a cDNA fragment (BM093886) and an RT-PCR product (Table S1 and Table S2), respectively. Results showed that steady state transcript levels of *F3H* and *ANS* were comparable among the soybean lines (Figure 2C). The probe for *DFR* was an RT-PCR product generated from immature flowers using primers DFR1F and DFR1R (Table S1 and Table S2). It encoded a protein named *DFR2* that showed 81% amino acid identity to *DFR1* (AF167556). The steady state *DFR2* transcript level was highest in wild-type Harosoy (*W4*) and the purple petal sectors of

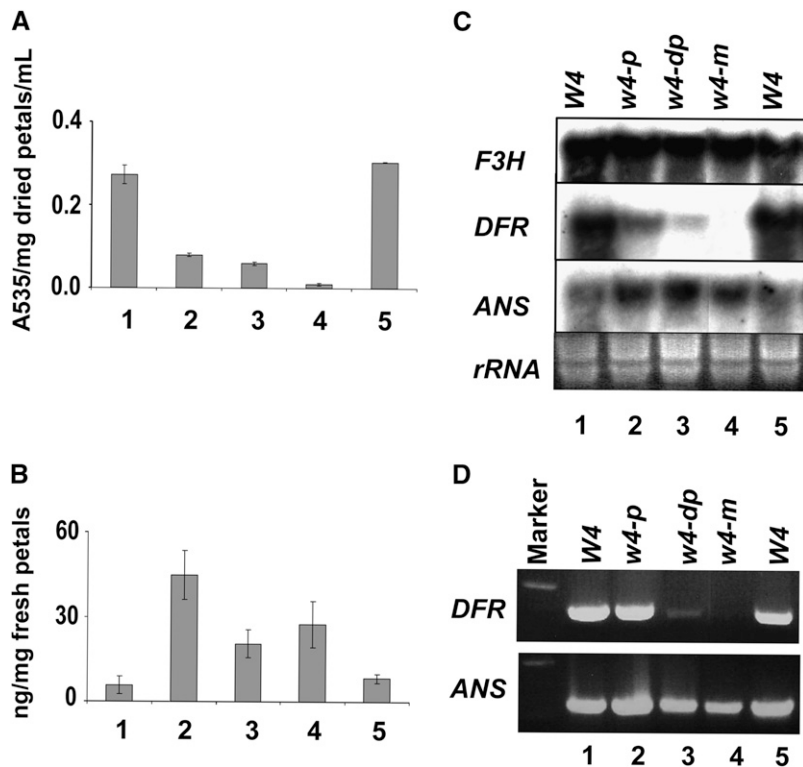


FIGURE 2.—Biochemical and molecular analyses of lines carrying individual *w4* alleles. (A) Contents of total anthocyanins in five immature petal samples from four soybean lines: 1, Harosoy (purple; *W4*); 2, T369 (pale; *w4-p*); 3, T321 (dilute purple; *w4-dp*); 4, T322w (white sectors of variegated petals; *w4-m*); and 5, T322p (purple section of variegated petals; *W4*). (B) Myricetin contents in five petal samples. (C) The transcript level of *DFR*, *ANS*, and *F3H* in five petal samples determined by Northern blot analysis. (D) RT-PCR products of *DFR* and *ANS* genes in five petal samples.

T322, reduced in T369 (*w4-p*) and T321 (*w4-dp*), and undetectable in the white petal sectors of T322 (Figure 2C). Similar results for steady state *DFR2* transcript levels were observed in RT-PCR analyses (Figure 2D). These data suggested that reduced levels of anthocyanin pigments in *w4* mutants (Figure 2A) were the results of lower *DFR2* expression levels.

**Characterization of T322 (*w4-m*) and its revertants suggested that *DFR2* is located in the *W4* locus:** *DFR2* was isolated from a BAC library (BHATTACHARYYA *et al.* 2005). BAC GS\_60E6 was selected for sequencing the entire *DFR2* gene (EF187612), which contains six exons and five introns (Figure 3A). The *DFR2* coding sequence predicted by GENSCAN (<http://genes.mit.edu/GENSCAN.html>) was 1065 bp long. The deduced *DFR2* polypeptide contained 354 amino acids with 82% identity to *DFR1* (AF167556) (Figure S3).

To determine whether *DFR2* is the *W4* gene and whether insertion of an active, class II transposable element in *DFR2* gave rise to the *w4-m* allele, cultivar Williams (*W4*), T322 (*w4-m*), and revertants of T322 [T321 (*w4-dp*), T369 (*w4-p*), and T325 (*W4*)] (Table 1) were studied for organization of *DFR2*. *DFR2* contains a *HindIII* restriction site in exon II, which divides the *DFR2* gene into two halves, 5' and 3' ends (Figure 3A). *DFR5'* and *DFR3'* cDNA probes hybridizing these 5' and 3' ends were prepared and used in Southern blot analyses (Figure 3A; Table S2). As expected, the ~5.5-kb *EcoRI* fragment was detected by both probes in Williams (*W4*), T321 (*w4-dp*), T369 (*w4-p*), and T325 (*W4*); but in T322 (*w4-m*), the fragment was ~6.8 kb (Figure 3B),

suggesting the presence of an insertion in the *w4-m* allele. Most likely this insertion was excised in the germinal revertants T321 (*w4-dp*), T369 (*w4-p*), and T325 (*W4*).

In *HindIII-PstI* double digested DNA, probe *DFR5'* detected a 1.7-kb *HindIII* fragment in Williams (*W4*), T322 (*w4-m*), and T325 (*W4*) as expected (Figure 3A), but an ~2.6-kb fragment in T321 (*w4-dp*) and an ~1.5-kb fragment in T369 (*w4-p*), respectively (Figure 3C, left panel). Probe *DFR3'* detected an ~2.8-kb fragment in Williams (*W4*), T321 (*w4-dp*), T369 (*w4-p*), and T325 (*W4*), but an ~2.3-kb fragment in T322 (*w4-m*) (Figure 3C, right panel). These results suggested that in T322, the insertion is located in the *HindIII-PstI* *DFR2* fragment.

In *DFR2*, the ~1.7-kb *HindIII* fragment includes the promoter, exon I, a part of exon II, and an *EcoRI* site (Figure 3A). Since no *DFR2*-specific polymorphisms for *EcoRI*-digested DNA was observed among wild-type and T321 (*w4-dp*) or T369 (*w4-p*) lines (Figure 3B), the aberrations in these two mutants should reside in the ~1.2-kb *HindIII-EcoRI* fragment containing the upstream promoter (Figure 3A). These results showed that *dfR2* mutations were generated from insertions among the *w4* alleles, and therefore *W4* most likely encodes *DFR2* (Figure 3, B and C).

**The insertion in *DFR2* intron II is a CACTA-like element *Tgm9*:** Southern analyses suggested that an insertion was located between *DFR2* exon II and VI in T322 (*w4-m*) (Figure 3). We isolated a 1357-bp insertion in *DFR2* intron II, 438 bp downstream of the exon II/

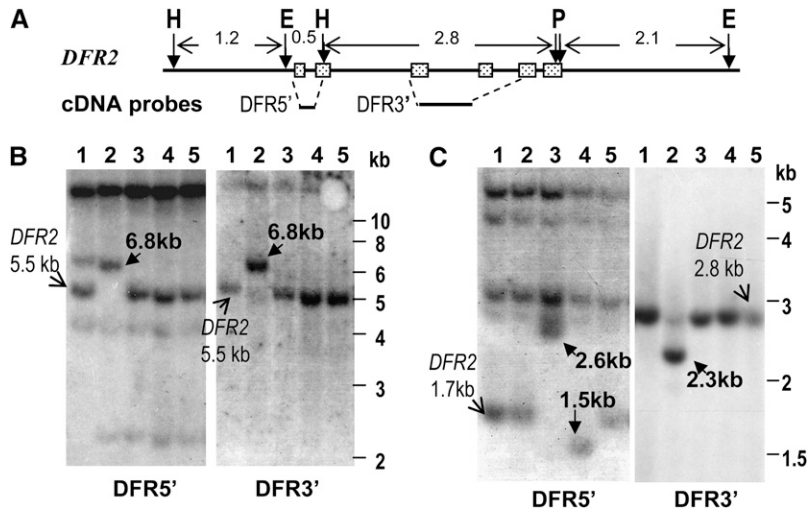


FIGURE 3.—Organization of *DFR2* among soybean lines carrying different *W4* alleles. (A) Schematic representation of the wild-type *DFR2* gene isolated from the BAC clone, GS\_60E6. E, *EcoRI*; H, *HindIII*; P, *PstI*. Approximate positions of the probes used in B and C are shown below the *DFR2* gene. (B) Organization of *EcoRI* digested DNA of five soybean lines. Probes are shown below each DNA blot. Slim arrows point to the expected restriction fragments from the wild-type *DFR2* gene. Boldface arrows point to the polymorphic fragments in *w4* mutants. Lanes 1, Williams (*W4*); 2, T322 (*w4-m*); 3, T321 (*w4-dp*); 4, T369 (*w4-p*); and 5, T325 (*W4*). (C) RFLP analyses of *HindIII*–*PstI* digested DNA of five soybean lines.

intron II junction (Figure 4A). The insertion harbors a *HindIII* site at the 3' end, which led to generation of an ~2.3-kb *HindIII*–*PstI* fragment for the *w4-m* allele when the DNA blot was hybridized with the *DFR3'* probe (Figure 3C).

The inserted element generated a 3-bp (AAT) target site duplication (TSD), similar to TSD generated by CACTA-type transposons (PEREIRA *et al.* 1986; RHODES and VODKIN 1988; NACKEN *et al.* 1991; INAGAKI *et al.* 1994) and contained structures similar to the 3' end of the CACTA elements. It carried a 30-bp terminal inverted repeat (TIR) starting with 5'-CACTA-3' similar to the ones in other soybean *Tgm* elements (Table S4) and a 700-bp highly repetitive region in the subterminal repeat (STR) region next to 3'-TIR (Figure 4). It does not contain other structures such as 5' end TIR and transposase gene(s), suggesting that it is a truncated version of a transposable element, most likely generated from an imperfect excision of the entire element. We named the entire element *Tgm9*.

To clone the entire *Tgm9*, we constructed and screened a genomic library carrying ~20 genome equivalents DNA prepared from T322 that showed high levels of both somatic excision and germinal reversion. Two nonoverlapping plaques, 16 and 25 carrying 5' and 3' ends of *Tgm9*, respectively, were sequenced (Figure S4 A). By conducting a long range PCR and then a sub PCR (Figure S4 B), a 19-bp (GTTTTGTTGATCATTTACA) missing *Tgm9* sequence between the two adjacent ends of clones 16 and 25 was obtained (Figure S4 A). *Tgm9* was 20,548 bp (GQ344503). It contained 5'- and 3'-TIR starting with 5'-CACTA-3', and transposase genes (Figure 4B).

The truncated *Tgm9* element was identical to the 3' end of *Tgm9* except for a novel 26-nt sequence (5'-ATTACGTACCATTTCAGTGAAATCACG-3'), which with its downstream 17-nt sequence (5'-TACCATTTCAGTGAAATC-3') formed two 20-bp tandem direct repeats (5'-ACGTACCATTTCAGTGAAATC-3') at the 5' end of

the truncated element (Figure 4B). We were able to PCR amplify the truncated element from T322. Therefore, the truncated element most likely arose from imprecise excision of the element. The novel 26-nt sequence was presumably generated through slipped mispairing accompanied by intragenic recombination and deletion as has been documented for generation of a direct repeat (TAVASSOLI *et al.* 1999).

*Tgm9* showed high sequence identity to *Tgmt\** isolated recently from the soybean *t\** allele (ZABALA and VODKIN 2008). Only 19 mismatches and 7 half mismatches (6 Rs and 1Y in *Tgmt\**) were found between the two elements (Table S3). The element was defined by an imperfectly inverted repeat starting with 5'-CACTA-3'. The 3'-STR end was highly structured and contained 12 stem-loop structures, each with a 7-bp motif (5'-AACCGTC-3') (ZABALA and VODKIN 2008). We observed that this 7-bp motif was located within a conserved 11-bp motif (5'-AACCGTCTTAR-3'). This conserved (80%) motif repeated 30 times as 15 tail-to-tail dimers in the 3'-STR region and 6 times as 3 tail-to-tail dimers in the 5'-STR region (Figure 4B).

**Alternate splicing generated transposase transcripts in *Tgm9*:** ZABALA and VODKIN (2008) identified 24 exons from the *Tgmt\** element. All these exons were found in *Tgm9* (Figure 4B, exons VI–XXVII) and their expression was detected by RT-PCR in T322. The exons contained two open reading frames (ORF), ORF1 and ORF2 (Figure 4B). By conducting rapid amplification of 5' complementary DNA ends (5'-RACE), we were able to identify three additional exons (exons I, II, and III) at the 5' end of the transcripts (Figure 4B). RT-PCR experiments revealed four types of transposase transcripts, t1–t4 (Figure 5).

The t1 and t2 transcripts contain ORF2 (Figures 4B and 5B). The 5' ends of these two transcripts were detected with a forward primer (P1) from exon I and a reverse primer (P4) from exon VI. The 5' end of t1 contained exons I, II, III, V, and VI, and that of t2

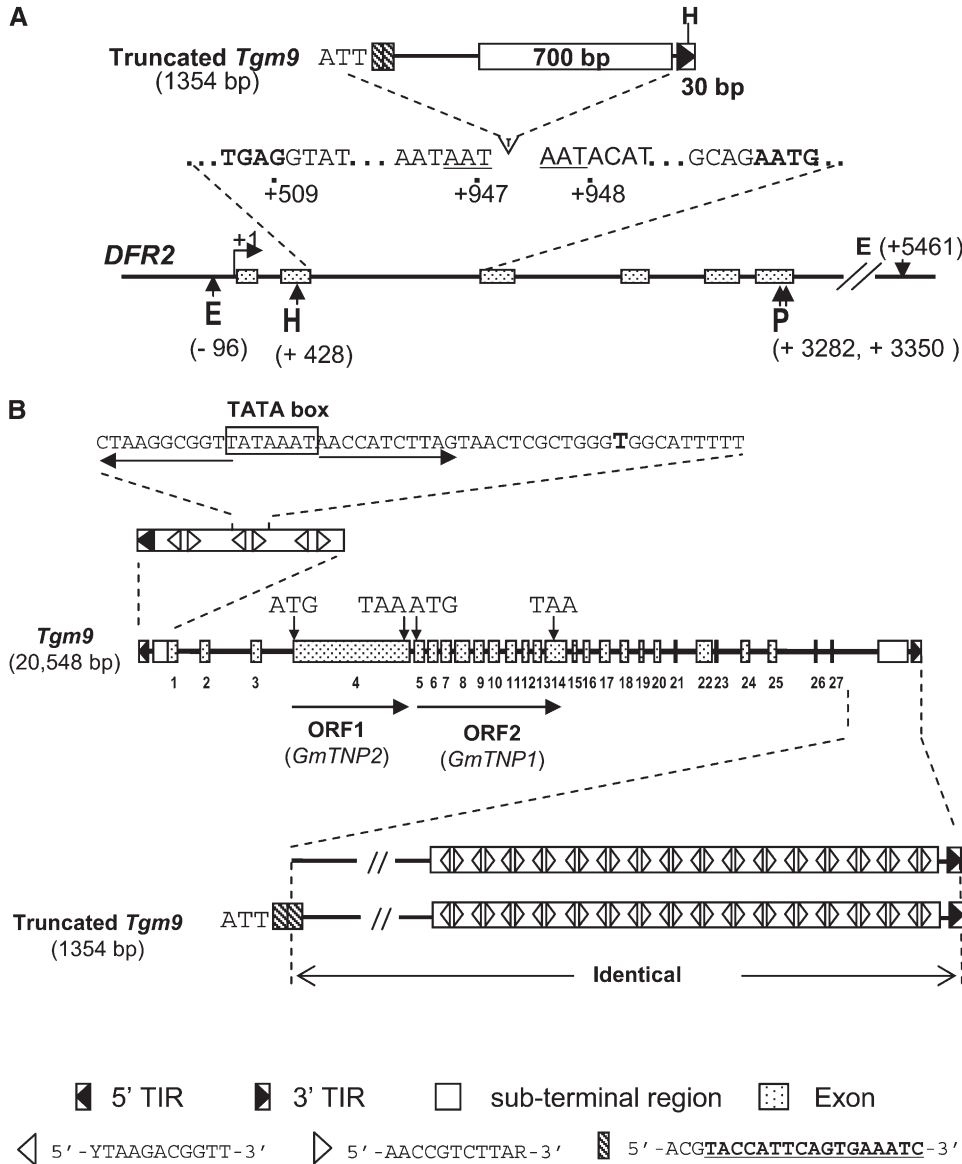


FIGURE 4.—Molecular characterization of the *Tgm9* element. (A) Schematic representation of the truncated *Tgm9* isolated from the *w4-m* allele. The positions of nucleotide and restriction sites relative to the transcription start site (TSS) (+1) in wild-type *DFR2* are shown. Nucleotides in bold-face type were from exons. E, *EcoRI*; H, *HindIII*; and P, *PstI*. (B) Schematic representation of the *Tgm9* element. *Tgm9* contains a 5'-TIR, 3'-TIR, and 27 exons (marked from 1 to 27) and two ORFs. ORF1 encodes putative transposase GmTNP2. ORF2 encodes putative transposase GmTNP1. A tail-to-tail dimer consisting of two inverted 11-bp motifs repeated 15 times at the 3' subterminal region (STR) and 3 times at the 5'-STR. A putative TATA box was identified from the second dimer of the 5' end STR. Truncated *Tgm9* element is identical to the 3' end of the full length *Tgm9* element with the exception of the 26 nucleotides (5'-ATTACGTACCATTTCAGTGAAATCAGC-3') in its 5' end that are absent in *Tgm9*. As a result, two 20-bp (ACGTACCATTTCAGTGAAATC) tandem repeats (box filled in with slashes) were identified from the 5' end of the truncated element.

included exons I, II, V, and VI (Figure 5, A and B). The 5'-UTR and ORF2 were identified in exons I–III and exons V–XIV, respectively. Exon IV containing ORF1 was spliced out in both t1 and t2. ORF2, starting at nt 9455 (exon V) and stopping at nt 12,546 (exon XIV), encoded a 755-aa polypeptide containing pfam03017 domain, which belonged to TNP1-like transposase 23 (<http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>) (Figure 6B). The deduced polypeptide was named GmTNP1. The N terminus of GmTNP1 shared 24% identity with transposase TNP1 in *Antirrhinum majus Tam1* (NACKEN *et al.* 1991) but no similarity with transposase TNPA in maize *En/Spm* (PEREIRA *et al.* 1986; MASSON *et al.* 1989).

The 5' ends of t3 and t4 transcripts contained exons I–IV and exons I, II, and IV, respectively (Figure 5). The first three exons constitute the 5'-

UTR of ORF1. ORF1, starting at nt 6127 and stopping at nt 9316 (exon IV) (Figure 4B), encoded a 1063-aa polypeptide with a conserved domain, pfam02992 (transposase\_21, transposase familyTNP2) found in TNP2 of *Tam1* (NACKEN *et al.* 1991) and TNPD of *En/Spm* (PEREIRA *et al.* 1986; MASSON *et al.* 1989) (Figure 6A). The deduced polypeptide was named GmTNP2, which shared 32 and 46% identities with TNP2 and TNPD, respectively.

**W4 encodes DFR2:** To determine whether the variegated flower phenotype is caused by excision of *Tgm9* from *DFR2*, we investigated >320 progenies of 21 families descended from a single T322 progenitor for hypocotyls and flower colors in greenhouse (Figure 7A). Nine families carried at least some progenies that were either germinal (purple hypocotyls and flowers) or somatic revertants (variegated flowers and purple sectors on hypocotyls). Six other families produced at least

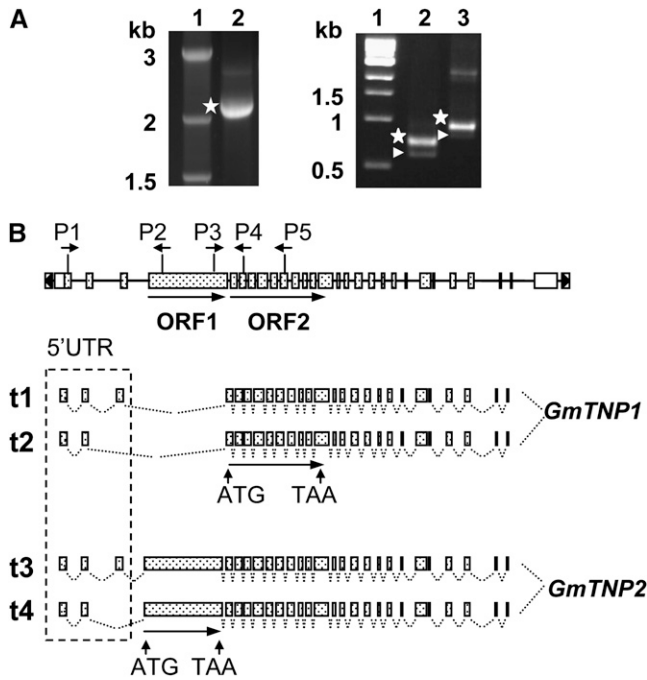


FIGURE 5.—Expression and alternative splicing of transposase genes in *Tgm9*. (A) Expression of transposase genes detected in T322 petals by RT-PCR. Twenty-seven exons of *Tgm9* were amplified in several RT-PCR experiments as segments and sequenced. For example, cDNA fragment (marked as star) amplified by primers P3 and P5 is shown in lane 2 of the left panel. It covered the region from ORF1 to ORF2. The 5'-UTRs of ORF1 and ORF2 were amplified with primers P1 and P2 (lane 2) or P1 and P4 (lane 3), respectively and are shown in the right panel. Positions of the primers are shown in B. Each primer combination produced two products. Products containing exon III were marked with stars or those without were with arrowheads. Lane 1, marker. (B) Schematic representation of *GmTNP1* and *GmTNP2* transcripts produced by alternative splicing. Four types of transcripts (t1–t4) with or without exon III or IV were detected. Transcripts t1 and t2 carrying no exon IV, amplified by primers P1 and P4 in A, encode *GmTNP1*; and t3 and t4 carrying exon IV, amplified by primers P1 and P2 in A, encode *GmTNP2*. Two ORFs and positions of their start/stop codons and 5'-UTR are shown.

some progenies that showed somatic excisions. The average rates of germinal reversion and somatic excisions were 4 and 25%, respectively (Figure 7B), which were comparable to earlier estimates (GROOSE *et al.* 1990). A larger proportion (>70%) of the progenies had only white flowers. Imprecise excision of *Tgm9* leading to truncated element (Figure 4B) in the target site could be one of the reasons for generation of high proportions of progenies with white flowers (Figure 7A).

We sequenced *Tgm9* insertion sites of independent germinal revertants with purple flowers and observed distinct footprints among the independent germinal revertants (Figure S5). These results confirmed that excision of *Tgm9* from the *DFR2* intron II resulted in the expression of *DFR2*, and thereby, gain of purple flower phenotype. Therefore, *W4* encodes *DFR2* and somatic

excision of the element results in variegated flower phenotype.

*w4-dp* and *w4-p* alleles were generated from reinsertion of *Tgm9* into the *DFR2* promoter: T321 (*w4-dp*) and T369 (*w4-p*) mutants were descended from T322 (*w4-m*). Sequencing of the *Tgm9* insertion site confirmed that *Tgm9* was excised from *DFR2* in both mutants and left behind 4- and 0 (precise excision)-bp footprints in T321 and T369, respectively (Figure S5). The 944-bp insertion (EU068463) in T321 was amplified using primers DFR4S and DFR4R (Table S1). It is identical to the 5' end of *Tgm9*. Two nucleotides (C and T) at the 3' end of insertion site (–1044) were deleted (Figure 8). We failed to PCR amplify the entire insertion in T369. Its 3' end (381 bp), PCR amplified with primers Tn3'1S and DFR4R (Table S1), was identical to the 3' end of *Tgm9* and located upstream of the –1034th nt of the *DFR2* promoter.

The insertion sites in the *w4-dp* and *w4-p* alleles were only 9 bp apart (Figure 8). The promoter regions between the insertion sites and the transcription start site (TSS) were PCR amplified and sequenced from T321 (*w4-dp*), T369 (*w4-p*), and T322 (*w4-m*). No rearrangements in this region occurred in the mutants (Figure S6). Therefore, the region upstream of *Tgm9* insertion sites is important for full expression of *DFR2*. The upstream promoter regions of structural anthocyanin biosynthesis genes contained *cis* regulatory elements that affect pigmentation patterns or intensity (COEN *et al.* 1986; ALMEIDA *et al.* 1989; LISTER *et al.* 1993). Putative *cis*-regulatory elements CCAAT motif (GELINAS *et al.* 1985) and E-box (CACGTG) (EPHRUSSI *et al.* 1985) are located upstream of *Tgm9* insertion sites in T321 (*w4-dp*) and T369 (*w4-p*) (Figure 8), which were moved away from the TSS in both mutants, presumably resulting in reduced expression of *DFR2* (Figure 2C).

***Tgm9* is a low copy number element:** CACTA elements usually have relatively low copy numbers (<100 copies) (KUNZE *et al.* 1997). An earlier study showed that the soybean genome contained 30–42 copies of the *Tgm*-like elements (RHODES and VODKIN 1988). The genomic DNA from three NILs, T322 (*w4-m*), T321 (*w4-dp*), and T325 (*W4*), were digested with *EcoRI* or double digested with *HindIII* and *PstI*, and DNA blots were hybridized to the 3' end of *Tgm9*. More than 10 copies of the *Tgm9*-like sequences were detected (Figure 9). T325 was isolated as a germinal revertant with purple flowers from T322. *HindIII* and *PstI* digested DNA showed the excision of *Tgm9* from the *DFR2* intron II and reinsertion into a new locus (Figure 9).

The recently available soybean genome sequence (<http://www.phytozome.org>) was searched for *Tgm9* 5' end (400 bp), 3' end (700 bp), *GmTNP1*, and *GmTNP2* sequences. The 5' end showed similarities to 32 sequences. The 3' end and *GmTNP1* showed similarities to ~100 sequences of the soybean genome. At least

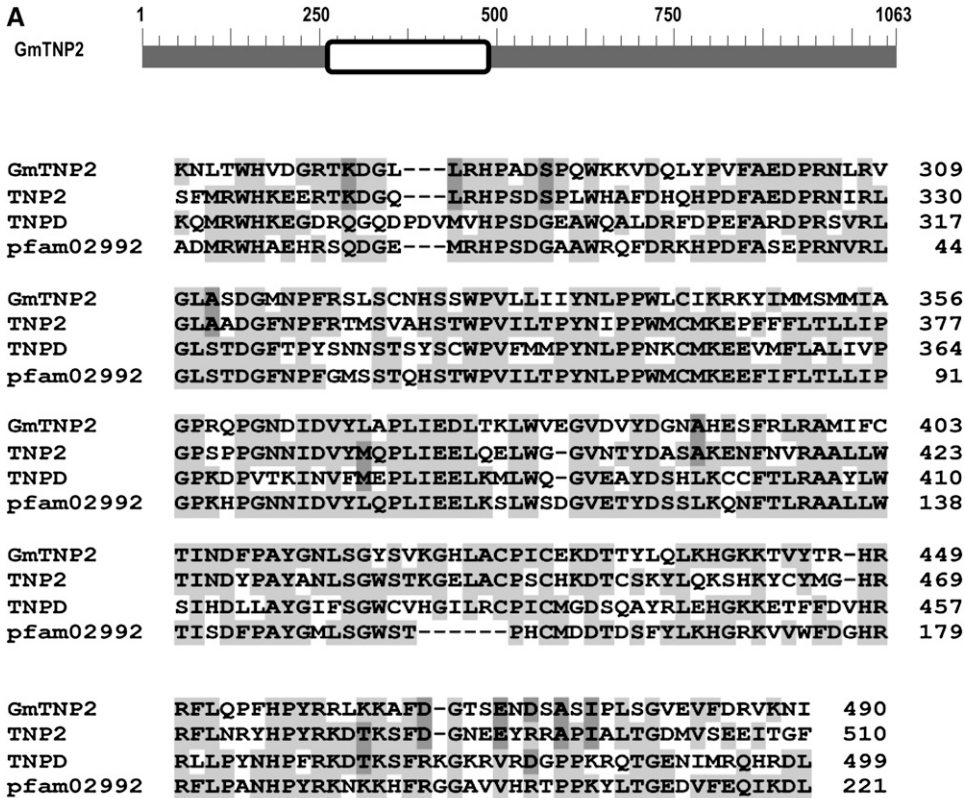
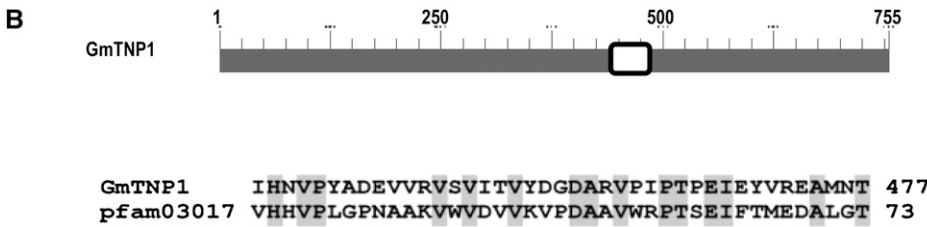


FIGURE 6.—Conserved domains in GmTNP1 and GmTNP2. (A) Schematic representation of GmTNP2. The conserved domain is boxed. The conserved sequences of the domain among GmTNP2, *Tam1* TNP2, *En/Spm* TNPD, and TNP2\_like domain pfam02992 (transposase\_21) are shown. (B) Schematic representation of GmTNP1. The conserved domain is boxed. The conserved sequences of the domain between GmTNP1 and TNP1-like domain pfam03017 super family (transposase\_23, TNP1/*En/Spm*) are shown. Conserved amino acids among different proteins are shaded.



1500 bp *GmTNP2* sequences showed similarity to 1000 sequences of the genome. This suggested that a TNP2-like domain could be conserved among different CACTA elements such as *Tgm5* (RHODES and VODKIN 1988) or functionally related distant proteins. Among the *Tgm9*-like sequences, one localized to scaffold\_57

from nt 95,650 to 13,598 is 99% identical to *Tgm9*. We named this sequence *Tgm10*. Compared to *Tgm9*, *Tgm10* is truncated for the first ~4100-bp sequence, contains a gap in its 5' end, and a 1049-bp insert in exon XXIII (Figure S7). *Tgm\**, *Tgm 9*, and *Tgm10* could be variants from a progenitor element or

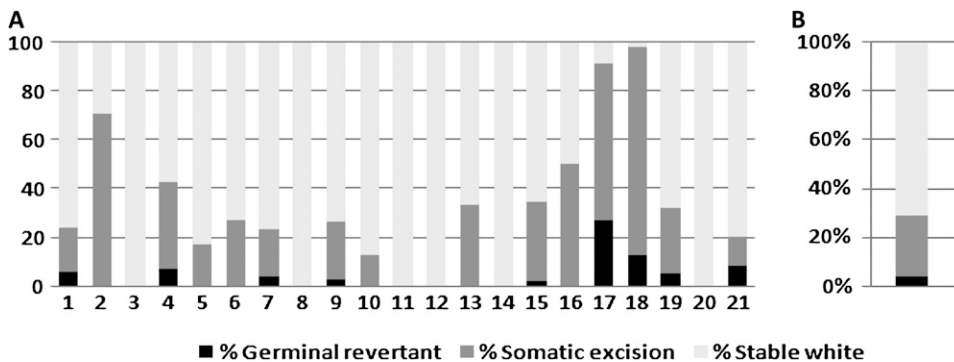


FIGURE 7.—*Tgm9* is a highly active endogenous transposable element. (A) Investigation of the rates (%) of germinal reversion and somatic excision for flower colors in 21 families generated from a single T322 plant. Germinal reversion produced only purple flowers. Somatic excision produced variegated flowers. Stable white plants produce only white flowers. (B) Percentages of progenies showing germinal reversion with only purple flowers, somatic reversion showing variegated petals and only stable white flowers among all 21 families shown in A.



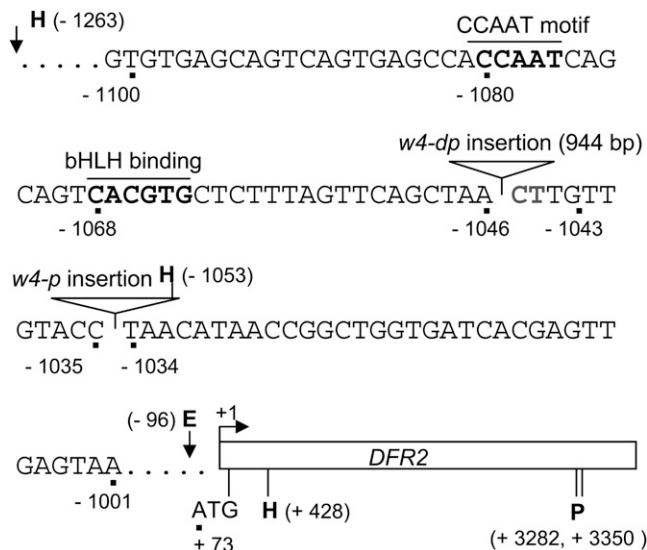


FIGURE 8.—Characterization of the *w4-dp* and *w4-p* alleles arisen following excision of *Tgm9* from the *DFR2* intron II. The positions of nucleotide or restriction sites relative to TSS (+1) are shown. Solid triangles indicate the location of insertions in *w4-dp* and *w4-p* alleles. E, *EcoRI*; H, *HindIII*; and P, *PstI*.

alternatively, highly active *Tgm9* could be a progenitor of *Tgmt\** and *Tgm10*.

## DISCUSSION

In soybean, the *w4-m* allele regulates variegated flower color in petals and purple sectors on stems or hypocotyls. By applying biochemical and molecular approaches, we have established that somatic excision of a CACTA-type transposable element *Tgm9* from *DFR2* encoding dihydroflavonol-4-reductase results in variegated flowers in mutable T322 line carrying the *w4-m* allele. *Tgm9* is ~20.5 kb long and a member of the CACTA super family of transposons (PEREIRA *et al.* 1986; RHODES and VODKIN 1988; NACKEN *et al.* 1991; INAGAKI *et al.* 1994). It generates 3-bp target-site duplication upon insertion. Its 5' and 3' ends carry imperfect terminal inverted repeats (TIRs) flanking the conserved CACTA sequence. Subterminal regions are highly structured and contain multiple copies of putative transposase binding motif (AACCGTCTTAR) (Figure 4) (GIERL *et al.* 1988). It excises at a high frequency (Figure 7). Excision of *Tgm9* generated 8- to 5-bp footprints (Figure S5), which are comparable to the ones created by other CACTA elements such as petunia *PstI* (SNOWDEN and NAPOLI 1998).

The excision mechanism in *Tgm9* could be similar to one considered for *En/Spm* (GIERL *et al.* 1988, 1989; FREY *et al.* 1990). Through alternative splicing, *Tgm9* produces two distinct transposases, GmTNP1 (755 aa, *Tam1* TNP1-like transposase) and GmTNP2 (1063 aa, *Tam1* TNP2-like transposase) (Figures 4B, 5, and 6). Organization of GmTNP2 and GmTNP1 is comparable

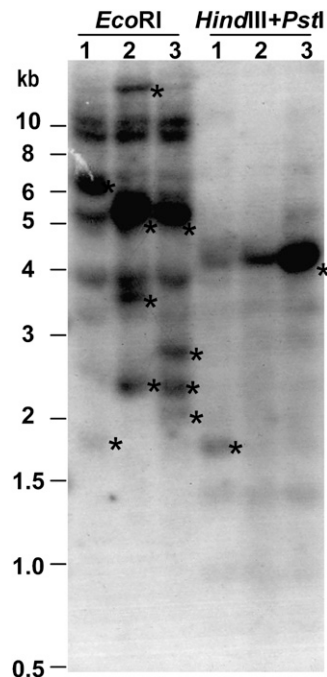


FIGURE 9.—Organization of *Tgm9* among soybean lines that vary for the *W4* alleles. The probe is the 3' end of *Tgm9*. Soybean lines, their *W4* alleles, and restriction enzymes used were labeled above individual lanes. Lanes 1, T322 (*w4-m*); 2, T321 (*w4-dp*); and 3, T325 (*W4*). T321 and T325 were isolated as intermediate or full revertant lines, respectively, from T322. Polymorphic bands are shown with \*. The *EcoRI*-specific polymorphisms among three lines were arisen most likely due to cytosine methylation in some of the *EcoRI* sites. The strong ~4.5-kb *HindIII-PstI* band in T325 indicated additional copy resulting from reinsertion of *Tgm9* into a new locus.

to the one observed for transposases TNPA and TNPD in the maize *En/Spm* element (PEREIRA *et al.* 1986; MASSON *et al.* 1989). GmTNP1 is presumably a DNA-binding protein like TNPA, recognizing and binding to the short repetitive motif of the subterminal regions (GIERL *et al.* 1988). GmTNP2 most likely is an endonuclease-like TNPD (GIERL *et al.* 1989; FREY *et al.* 1990). It binds to GmTNP1, interacts with *Tgm9* TIRs, pulls the two ends of the element together to form a loop, and excises the element from its insertion site.

The maize *En/Spm* element preferentially transposes to linked loci (PETERSON 1970; NOWICK and PETERSON 1981). Similarly, *Tgm9* transposed into the *DFR2* promoter (Figures 3 and 8). However, with the exception of the mutations induced in the *DFR2* promoter (*e.g.*, *w4-dp* and *w4-p*) (Figures 3 and 8), mutations identified in *Tgm9*-tagging experiments were mapped to unlinked loci (PALMER *et al.* 1989, 2008a,b; KATO and PALMER 2003, 2004; XU and PALMER 2005b).

*Tgm9* showed high identity to the *Tgmt\** element (EU190440, 20,544 bp) isolated from the soybean *t\** allele (ZABALA and VODKIN 2008) (Table S3). As shown here and earlier, *Tgm9* is an active element (PALMER *et al.* 1989; GROOSE *et al.* 1990) (Figure 7); whereas, *Tgmt\** at the soybean *t\** allele seems not to

be (ZABALA and VODKIN 2003, 2008). The transposase genes were silenced in line 37609 (*t\**) (ZABALA and VODKIN 2008). High similarity between *Tgmt\** and *Tgm9* suggested that *Tgm9* could be the progenitor element of *Tgmt\**. *Tgmt\** is comparable to the cryptic *spm* element from the maize *a-m2-8167B* allele that contained an intact *spm* element with no activity (MASSON *et al.* 1987; BANKS *et al.* 1988).

Like most CACTA elements, *Tgm9* is a low copy transposable element (RHODES and VODKIN 1988; KUNZE *et al.* 1997). Active low copy endogenous transposable elements have been considered useful tools in gene cloning and functional genomics studies (MAES *et al.* 1999; WALBOT 2000; RAMACHANDRAN and SUNDARESAN 2001). We expect that highly active *Tgm9* should facilitate functional genomics studies in soybean. Genetic data strongly suggested that mutations such as necrotic root (*rn*), male-sterility, and female sterility (*st8*) (PALMER *et al.* 2008a,b) most likely resulted from insertion of *Tgm9*. Except for two mutations in fertility genes, no reversions events have been observed among the mutants presumably tagged by *Tgm9* (R. PALMER, unpublished data).

Truncated *Tgm10* and fractured *Tgm9* in *w4-dp* and *w4-m* allele (EU068463 and EU068464; Figure S7) suggested existence of fractured *Tgm9* elements in the soybean genome. Fractured *Ac* (*fAc*) elements have been documented in maize (RALSTON *et al.* 1989; ZHANG AND PETERSON 1999). We hypothesize that the element is frequently fractured during transposition events and truncated *Tgm9* derivatives cause stable mutations. If our hypothesis is correct, then the element will be useful in creating stable mutations and cloning soybean genes through *Tgm9*-tagging experiments. To date, to our knowledge no active, endogenous transposable elements have been cloned from any legume species. Therefore, *Tgm9* is expected to expedite the genomics research in soybean, and thereby contribute significantly toward our understanding of the legume biology.

The authors thank R. W. Groose, University of Wyoming, Laramie, WY, for providing flower pictures; R. C. Shoemaker, United States Department of Agriculture Agricultural Research Service (USDA ARS) and Iowa State University, Ames, IA, for providing the EST clone Gm-c1086-2103; and M. P. Scott, USDA ARS and Iowa State University, for the guidance on the HPLC analysis. We also thank R. Takahashi, National Institute of Crop Science, Tsukuba, Japan and L. Vodkin, Department of Crop Sciences, University of Illinois, Champaign-Urbana, IL for critically reviewing an earlier version of the manuscript and Cathie Martin, John Innes Center, United Kingdom, for reviewing the manuscript. This is a joint contribution of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, Project No. 4403, and the USDA, Agricultural Research Service, Corn Insects and Crop Genetics Research Unit, and was supported by the Hatch Act and the State of Iowa. The mention of a trademark or proprietary product does not constitute a guarantee or warranty of the product by Iowa State University or the USDA, and the use of the name by Iowa State University or the USDA implies no approval of the product to the exclusion of others that may also be suitable.

## LITERATURE CITED

- ALMEIDA, J., R. CARPENTER, T. P. ROBBINS, C. MARTIN and E. S. COEN, 1989 Genetic interactions underlying flower color patterns in *Antirrhinum majus*. *Genes Dev.* **3**: 1758–1767.
- BANKS, J. A., P. MASSON and N. FEDOROFF, 1988 Molecular mechanisms in the developmental regulation of the maize Suppressor-mutator transposable element. *Genes Dev.* **2**: 1364–1380.
- BHATTACHARYYA, M. K., N. N. NARAYANAN, H. GAO, D. K. SANTRA, S. S. SALIMATH *et al.*, 2005 Identification of a large cluster of coiled coil-nucleotide binding site–leucine rich repeat-type genes from the *Rps1* region containing Phytophthora resistance genes in soybean. *Theor. Appl. Genet.* **111**: 75–86.
- BURBULIS, I. E., M. IACOBUCCI and B. W. SHIRLEY, 1996 A null mutation in the first enzyme of flavonoid biosynthesis does not affect male fertility in *Arabidopsis*. *Plant Cell* **8**: 1013–1025.
- COEN, E. S., R. CARPENTER and C. MARTIN, 1986 Transposable elements generate novel spatial patterns of gene expression in *Antirrhinum majus*. *Cell* **47**: 285–296.
- EPHRUSSI, A., G. M. CHURCH, S. TONEGAWA and W. GILBERT, 1985 B lineage-specific interactions of an immunoglobulin enhancer with cellular factors *in vivo*. *Science* **227**: 134–140.
- FASOULA, D. A., P. A. STEPHENS, C. D. NICKELL and L. O. VODKIN, 1995 Cosegregation of purple-throat flower color with dihydroflavonol reductase polymorphism in soybean. *Crop Sci.* **35**: 1028–1031.
- FREY, M., J. REINECKE, S. GRANT, H. SAEDLER and A. GIERL, 1990 Excision of the *En/Spm* transposable element of *Zea mays* requires two element-encoded proteins. *EMBO J.* **9**: 4037–4044.
- GELINAS, R., B. ENDLICH, C. PFEIFFER, M. YAGI and G. STAMATOYANNOPOULOS, 1985 G to A substitution in the distal CCAAT box of the A gamma-globin gene in Greek hereditary persistence of fetal haemoglobin. *Nature* **313**: 323–325.
- GIERL, A., S. LUTTICKE and H. SAEDLER, 1988 TnpA product encoded by the transposable element *En-1* of *Zea mays* is a DNA binding protein. *EMBO J.* **7**: 4045–4053.
- GIERL, A., H. SAEDLER and P. A. PETERSON, 1989 Maize transposable elements. *Annu. Rev. Genet.* **23**: 71–85.
- GROOSE, R. W., and R. G. PALMER, 1991 Gene action governing anthocyanin pigmentation in soybean. *J. Hered.* **82**: 498–501.
- GROOSE, R. W., H. D. WEIGELT and R. G. PALMER, 1988 Somatic analysis of an unstable mutation for anthocyanin pigmentation in soybean. *J. Hered.* **79**: 263–267.
- GROOSE, R. W., S. M. SCHULTE and R. G. PALMER, 1990 Germinal reversion of an unstable mutation for anthocyanin pigmentation in soybean. *Theor. Appl. Genet.* **79**: 161–167.
- HARBORNE, J. B., 1958 Spectral methods of characterizing anthocyanins. *Biochem. J.* **70**: 22–28.
- HARTWIG, E. E., and K. HINSON, 1962 Inheritance of flower color in soybeans. *Crop Sci.* **2**: 152–153.
- HOLTON, T. A., and E. C. CORNISH, 1995 Genetics and biochemistry of anthocyanin biosynthesis. *Plant Cell* **7**: 1071–1083.
- INAGAKI, Y., Y. HISATOMI, T. SUZUKI, K. KASAHARA and S. IIDA, 1994 Isolation of a suppressor-mutator/enhancer-like transposable element, *Tpm1*, from Japanese morning glory bearing variegated flowers. *Plant Cell* **6**: 375–383.
- KATO, K. K., and R. G. PALMER, 2003 Molecular mapping of the male-sterile, female-sterile mutant gene (*st8*) in soybean. *J. Hered.* **94**: 425–428.
- KATO, K. K., and R. G. PALMER, 2004 Molecular mapping of four ovule lethal mutants in soybean. *Theor. Appl. Genet.* **108**: 577–585.
- KEIM, P., T. OLSON and R. C. SHOEMAKER, 1988 A rapid protocol for isolating soybean DNA. *Soybean Genet. Newsl.* **15**: 150–152.
- KUNZE, R., H. SAEDLER and W. E. LÖNNIG, 1997 Plant transposable elements. *Adv. Bot. Res.* **27**: 331–370.
- LISTER, C., D. JACKSON and C. MARTIN, 1993 Transposon-induced inversion in *Antirrhinum* modifies *nivea* gene expression to give a novel flower color pattern under the control of *cycloidearadialis*. *Plant Cell* **5**: 1541–1553.
- MAES, T., P. DE KEUKELEIRE and T. GERATS, 1999 Plant tagology. *Trends Plant Sci.* **4**: 90–96.
- MASSON, P., R. SUROSKY, J. A. KINGSBURY and N. V. FEDOROFF, 1987 Genetic and molecular analysis of the *Spm*-dependent *a-m2* alleles of the maize *a* locus. *Genetics* **117**: 117–137.

- MASSON, P., G. RUTHERFORD, J. A. BANKS and N. FEDOROFF, 1989 Essential large transcripts of the maize *Spm* transposable element are generated by alternative splicing. *Cell* **58**: 755–765.
- NACKEN, W. K., R. PIOTROWIAK, H. SAEDLER and H. SOMMER, 1991 The transposable element *Tam1* from *Antirrhinum majus* shows structural homology to the maize transposon *En/Spm* and has no sequence specificity of insertion. *Mol. Gen. Genet.* **228**: 201–208.
- NOWICK, E. M., and P. A. PETERSON, 1981 Transposition of the enhancer controlling element system in maize. *Mol. Gen. Genet.* **183**: 440–448.
- NOZZOLILLO, C., 1973 A survey of anthocyanin pigments in seedling legumes. *Can. J. Bot.* **51**: 911–915.
- PALMER, R. G., and R. W. GROOSE, 1993 A new allele at the *w4* locus derived from the *w4-m* mutable allele in soybean. *J. Hered.* **84**: 297–300.
- PALMER, R. G., B. R. HEDGES, R. S. BENAVENTE and R. W. GROOSE, 1989 *w4*-mutable line in soybean. *Dev. Genet.* **10**: 542–551.
- PALMER, R. G., R. W. GROOSE, H. D. WEIGELT and J. E. MILLER, 1990 Registration of a genetic stock (*w4-m w4-m*) for unstable anthocyanin pigmentation in soybean. *Crop Sci.* **30**: 1376–1379.
- PALMER, R. G., T. W. PFEIFFER, G. R. BUSS and T. C. KILEN, 2004 Qualitative genetics, pp. 137–234 in *Soybeans: Improvement, Production, and Uses*, edited by J. E. SPECHT and H. R. BOERMA. American Society of Agronomy, Madison, WI.
- PALMER, R. G., D. SANDHU, K. CURRAN and M. K. BHATTACHARYYA, 2008a Molecular mapping of 36 soybean male-sterile, female-sterile mutants. *Theor. Appl. Genet.* **117**: 711–719.
- PALMER, R. G., L. ZHANG, Z. HUANG and M. XU, 2008b Allelism and molecular mapping of soybean necrotic root mutants. *Genome* **51**: 243–250.
- PEREIRA, A., H. CUYPERS, A. GIERL, Z. SCHWARZ-SOMMER and H. SAEDLER, 1986 Molecular analysis of the *En/Spm* transposable element system of *Zea mays*. *EMBO J.* **5**: 835–841.
- PETERSON, P. A., 1970 The *En* mutable system in maize III. Transposition associated with mutational events. *Theor. Appl. Genet.* **40**: 367–377.
- RALSTON, E., J. ENGLISH and H. K. DOONER, 1989 Chromosome-breaking structure in maize involving a fractured *Ac* element. *Proc. Natl. Acad. Sci. USA* **86**: 9451–9455.
- RAMACHANDRAN, S., and V. SUNDARESAN, 2001 Transposons as tools for functional genomics. *Plant Physiol. Biochem.* **39**: 243–252.
- RHODES, P. R., and L. O. VODKIN, 1988 Organization of the *Tgm* family of transposable elements in soybean. *Genetics* **120**: 597–604.
- SAMBROOK, J., E. F. FRITSCH and T. MANIATIS, 1989 *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- SNOWDEN, K. C., and C. A. NAPOLI, 1998 Psl: a novel *Spm*-like transposable element from *Petunia hybrida*. *Plant J.* **14**: 43–54.
- TAKAHASHI, R., S. M. GITHIRI, K. HATAYAMA, E. G. DUBOUZET, N. SHIMADA *et al.*, 2007 A single-base deletion in soybean flavonol synthase gene is associated with magenta flower color. *Plant Mol. Biol.* **63**: 125–135.
- TAVASSOLI, K., A. EIGEL and J. HORST, 1999 A deletion/insertion leading to the generation of a direct repeat as a result of slipped mispairing and intragenic recombination in the factor VIII gene. *Hum. Genet.* **104**: 435–437.
- WALBOT, V., 2000 Saturation mutagenesis using maize transposons. *Cult. Opin. Plant Biol.* **3**: 103–107.
- WEIGELT, H. D., R. G. PALMER and R. W. GROOSE, 1990 Origin of the *w4-m* allele. *Soybean Genet. Newsl.* **17**: 81–84.
- XU, M., and R. G. PALMER, 2005a Genetic analysis and molecular mapping of a pale flower allele at the *W4* locus in soybean. *Genome* **48**: 334–340.
- XU, M., and R. G. PALMER, 2005b Molecular mapping of *k2 Mdh1-n y20*, an unstable chromosomal region in soybean. [*Glycine max* (L.) Merr.] *Theor. Appl. Genet.* **111**: 1457–1465.
- ZABALA, G., and L. VODKIN, 2003 Cloning of the pleiotropic *T* locus in soybean and two recessive alleles that differentially affect structure and expression of the encoded flavonoid 3' hydroxylase. *Genetics* **163**: 295–309.
- ZABALA, G., and L. VODKIN, 2007 Novel exon combinations generated by alternative splicing of gene fragments mobilized by a CACTA transposon in *Glycine max*. *BMC Plant Biol.* **7**: 38.
- ZABALA, G., and L. VODKIN, 2008 A putative autonomous 20.5 kb CACTA transposon insertion in an F3'H allele identifies a new CACTA transposon subfamily in *Glycine max*. *BMC Plant Biol.* **8**: 124.
- ZABALA, G., and L. O. VODKIN, 2005 The *wp* mutation of *Glycine max* carries a gene-fragment-rich transposon of the CACTA superfamily. *Plant Cell* **17**: 2619–2632.
- ZHANG, J., and T. PETERSON, 1999 Genome rearrangements by non-linear transposons in maize. *Genetics* **153**: 1403–1410.

Communicating editor: E. J. RICHARDS

# GENETICS

Supporting Information

<http://www.genetics.org/cgi/data/genetics.109.107904/DC1/1>

**Excision of an Active CACTA-Like Transposable Element  
From *DFR2* Causes Variegated Flowers in Soybean  
[*Glycine max* (L.) Merr.]**

**Min Xu, Hargeet K. Brar, Sehiza Grosic, Reid G. Palmer  
and Madan K. Bhattacharyya**

Copyright © 2009 by the Genetics Society of America  
DOI: 10.1534/genetics.109.107904

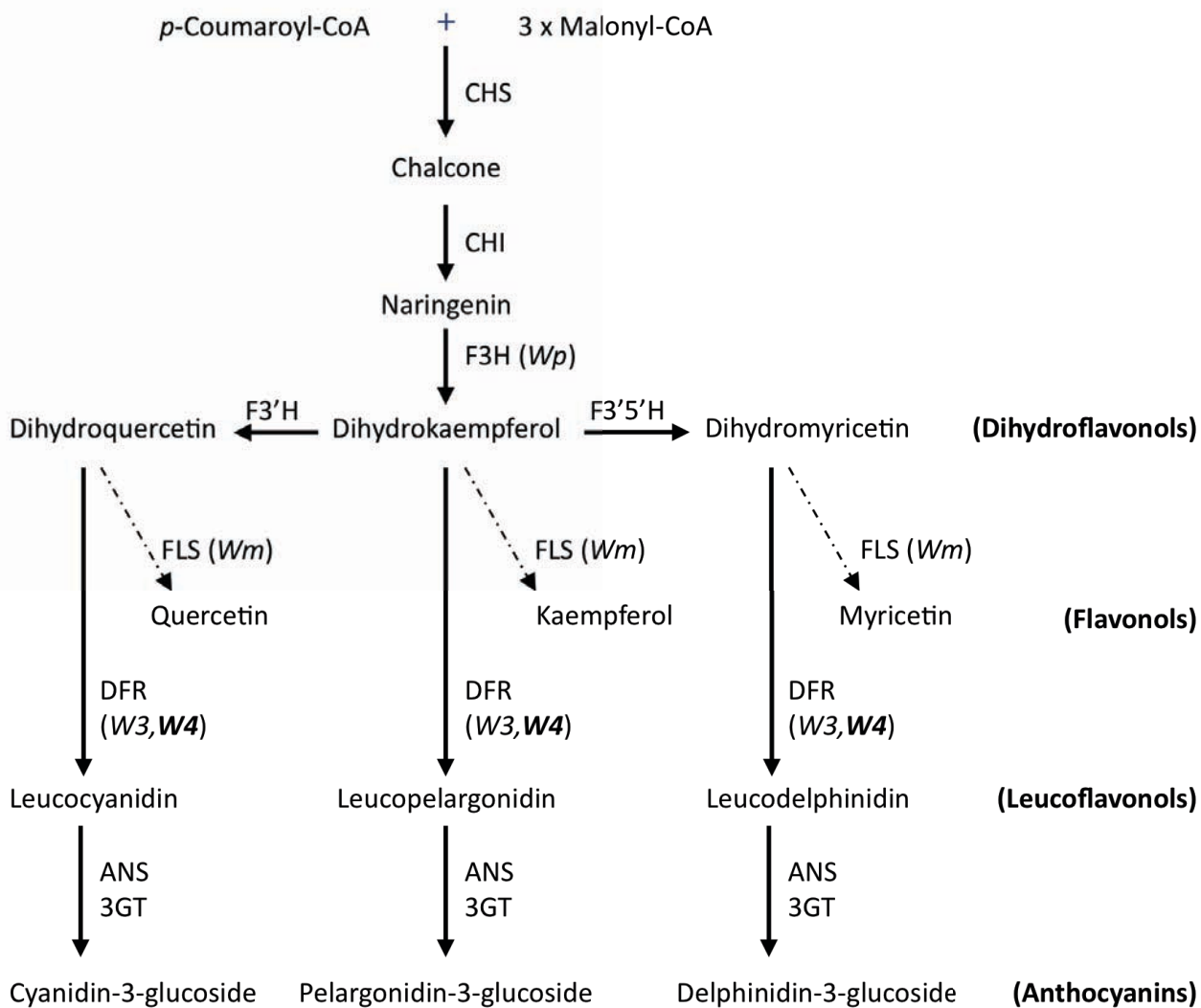


FIGURE S1.—Anthocyanin biosynthesis pathway involved in flower color development. CHS, Chalcone Synthase; CHI, Chalcone Isomerase; F3H, Flavanone 3-Hydroxylase; F3'5'H, Flavonoid 3'5'-Hydroxylase; F3'H, Flavonoid 3'-Hydroxylase; DFR, Dihydroflavonol-4-Reductase; ANS, Anthocyanidin Synthase; 3GT, 3- Glucose Transferase; FLS, Flavonol Synthase

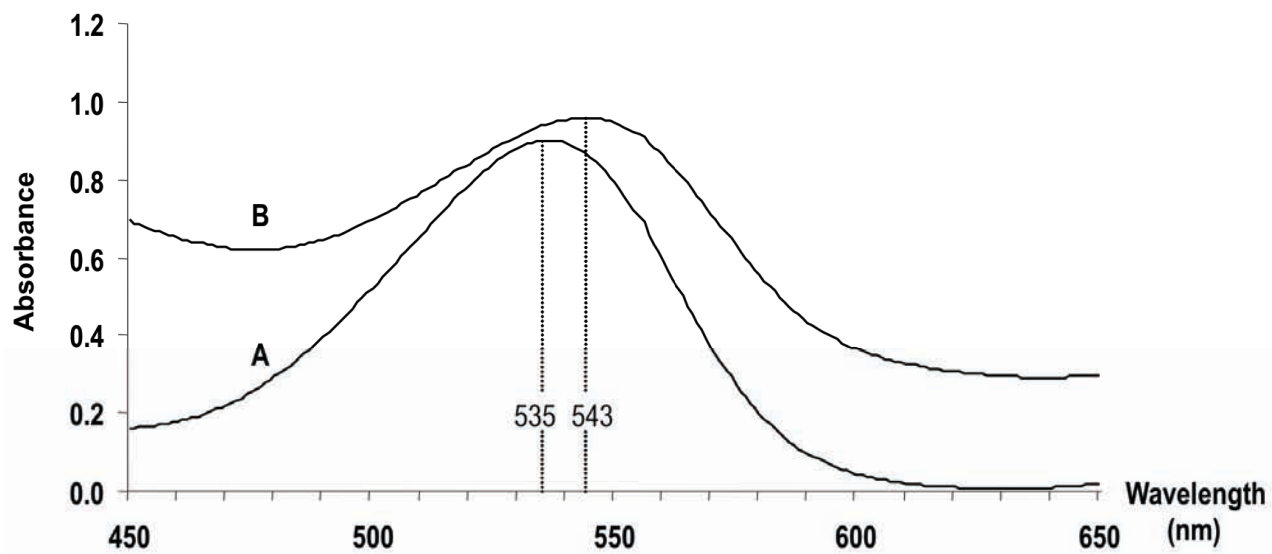


FIGURE S2.—Absorption spectra of anthocyanins extracted from immature flowers of the cultivar Harosoy. (A) Absorption spectrum of anthocyanin extracts in methanol-HCl, the absorption peak is at 535 nm. (B) Absorption spectrum of anthocyanin aglycones hydrolyzed by boiling 30 minutes, the absorption peak shifted 8 nm to 543 nm.

```

DFR2      MGSSSASESVCVTGASGFIGSWLMRLIERGYTVRATVVRDPANMKKVKHLVLPGAKTKL 60
DFR1      MG--SASESVCVTGASGFIGSWLMRLIERGYTVRATVVRDPVNMKKVKHLVLPGAKSKL 58
          ** *****;*****;.*****;*****;*****;*****;*****;*****;*****;*****

DFR2      SLWKADLAQEGSFDEAIKGC TGVFHVATPMDFDSKDPENEVIKPTINGLLDIMKACVKAK 120
DFR1      SLWKADLAEEGSFDEAIKGC TGVFHVATPMDFESKDPENEVIKPTINGVLDIMKACLKAK 118
          *****;.*****;.*****;.*****;.*****;.*****;.*****;.*****;*****;*****

DFR2      TVRRLVFTSSAGTVDVTEHPNPVIDENCWSDVDFCTRVKMTGWMYFVSKTLAEQEAWKYA 180
DFR1      TVRRLIFTSSAGTLNVIERQKPVFDDTCWSDVEFCRRVKMTGWMYFVSKTLAEKEAWKFA 178
          *****;.*****;.*****;.*****;.*****;.*****;.*****;.*****;.*****;*****

DFR2      KEHNIDFISVIPPLVVGPFLLMPTMPPSLITALSLITGNESHYHIIKQGQFVHLDDLCLGH 240
DFR1      KEQGLDFITIIIPPLVVGPFLLMPTMPPSLITALSPITGNEDHYSIIKQGQFVHLDDLCLAH 238
          **;.*****;.*****;.*****;.*****;.*****;.*****;.*****;.*****;*****

DFR2      IFVFNPKAEGRYICCSHEATIHDIAKLLNQKYPEYNVLTkfkNIPDELDIKfSSKkIT 300
DFR1      IFLFEEPEVEGRYICsACDATIHDIaklINQKYPEYKVPtKfKNIPDQLelVRfSSKkIT 298
          **;.*****;.*****;.*****;.*****;.*****;.*****;.*****;.*****;*****

DFR2      DLGFKFKYSLedMFTGAVETCREKGLLPKPeETTVNNeLLPKPAETTVNdtMQk 354
DFR1      DLGFKFKYSLedMYtGAIDtCRDKGLLPKPAEK---GLFtKPGETpVN-AMhk 347
          *****;.*****;.*****;.*****;.*****;.*****;.*****;.*****;.*****;*****

```

?

FIGURE S3.—Alignment of DFR1 with DFR2. “\*” represents identical residues; “:” means conserved substitutions between similar residues; “.” indicates the semi-conserved substitutions between similar residues.

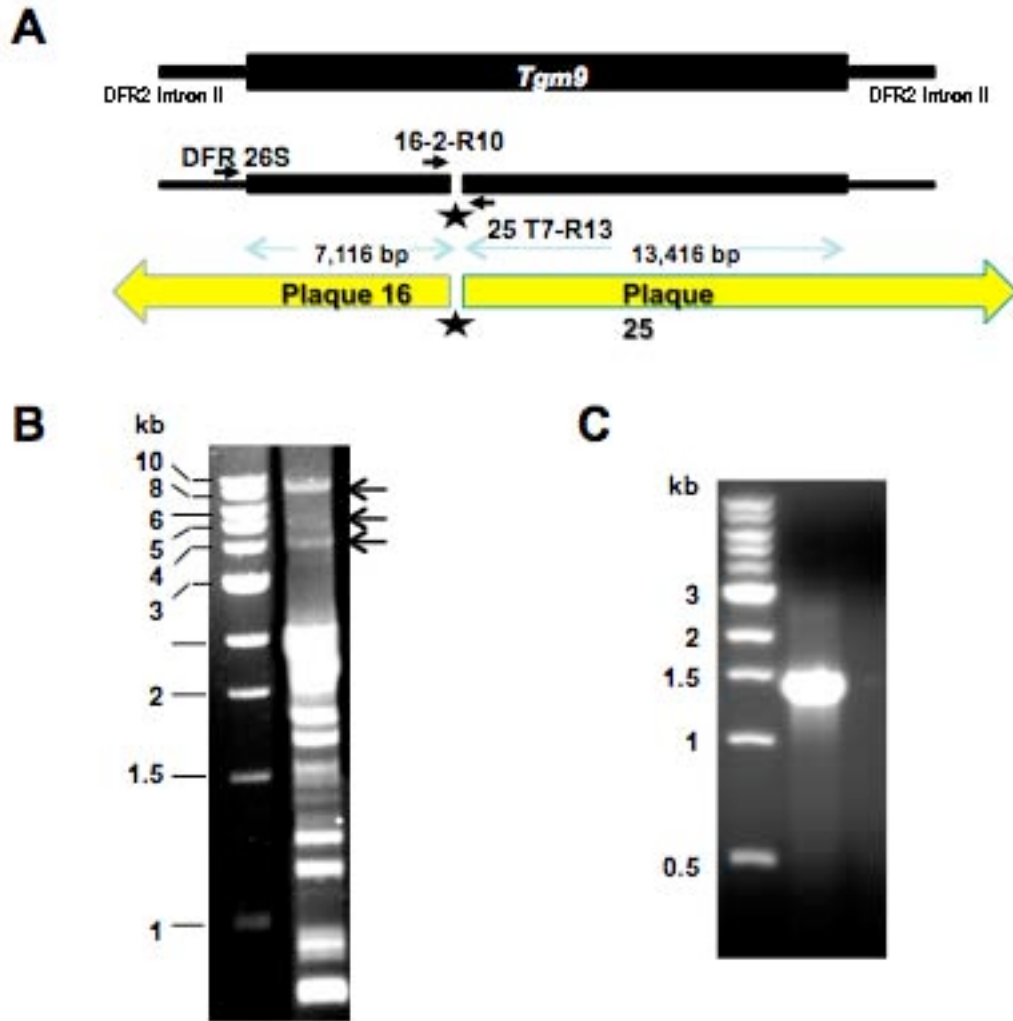



FIGURE S4.—PCR and sequencing strategies in obtaining a missing sequence of *Tgm9*. (A) Diagrammatic presentation of plaques 16 and 25 that were used to obtain most part of the *Tgm9* sequence. (B) Long-range (LR) PCR using primers DFR26S (5'-CAAGGACCCTGAGGTATGT-TGATCAT-3') and 25 T7-R13 (5'-CCCACACAAACGTATTTCTCGAAC-3') was applied to amplify the fragments including *DFR2* intron II and part of *Tgm9* as shown in (A). (C) Sub-PCR of the LR PCR products (shown by arrows in B). Sub-PCR product obtained by using primers 16-2-R10 (5'-GGTTCCTGGCGCTTCCAGTGAAG-3') and 25T7-R13 was used to obtain the sequence of the gap region shown by a black star in (A).



***Tgmw4m***



T322	CGTATCATATTTATATATTTTATAGGTAATAAT	<u>AA</u> TACATGAATGCTTATATTTTTT
58-1	CGTATCATATTTATATATTTTATAGGTAATAAT	ACATGAATGCTTATATTTTTT
58-18	CGTATCATATTTATATATTTTATAGGTAATA	<b>AA</b> TACATGAATGCTTATATTTTTT
58-9	CGTATCATATTTATATATTTTATAGGTAATAAT	<b>AT</b> ACATGAATGCTTATATTTTTT
58-4	CGTATCATATTTATATATTTTATAGGTAATAAT	<b>T AA</b> TACATGAATGCTTATATTTTTT
58-6	CGTATCATATTTATATATTTTATAGGTAATAAT	<b>AT</b> ACATGAATGCTTATATTTTTT
58-13	CGTATCATATTTATATATTTTATA	TACATGAATGCTTATATTTTTT
58-21	CGTATCATATTTATATATTTTATAGGTAATAAT	<b>T AA</b> TACATGAATGCTTATATTTTTT
58-7	CGTATCATATTTATATATTTTATAGGTAATAAT	<b>AT AA</b> TACATGAATGCTTATATTTTTT
58-5	CGTATCATATTTATATATTTTATAGGTAATAAT	<b>AT</b> ACATGAATGCTTATATTTTTT
T321	CGTATCATATTTATATATTTTATAGGTAATAAT	<b>AT AA</b> TACATGAATGCTTATATTTTTT
T369	CGTATCATATTTATATATTTTATAGGTAATAAT	CATGAATGCTTATATTTTTT
WT	CGTATCATATTTATATATTTTATAGGTAATAAT	ACATGAATGCTTATATTTTTT
	*****	*****

FIGURE S5.—Unique footprints left behind by *Tgm9* during germinal reversion. Germinal revertants from nine families identified in Figure 7a and two intermediate germinal revertants T321 (*w4-dp*) and T369 (*w4-p*) were selected for determining footprints left behind by *Tgm9* in *DFR2* intron II through PCR by compare to the wild-type *DFR2* (WT) from cv. Williams 82. Nucleotides representing the target site duplication are underlined. Footprint nucleotides left by *Tgm9* germinal excision are in bold font.

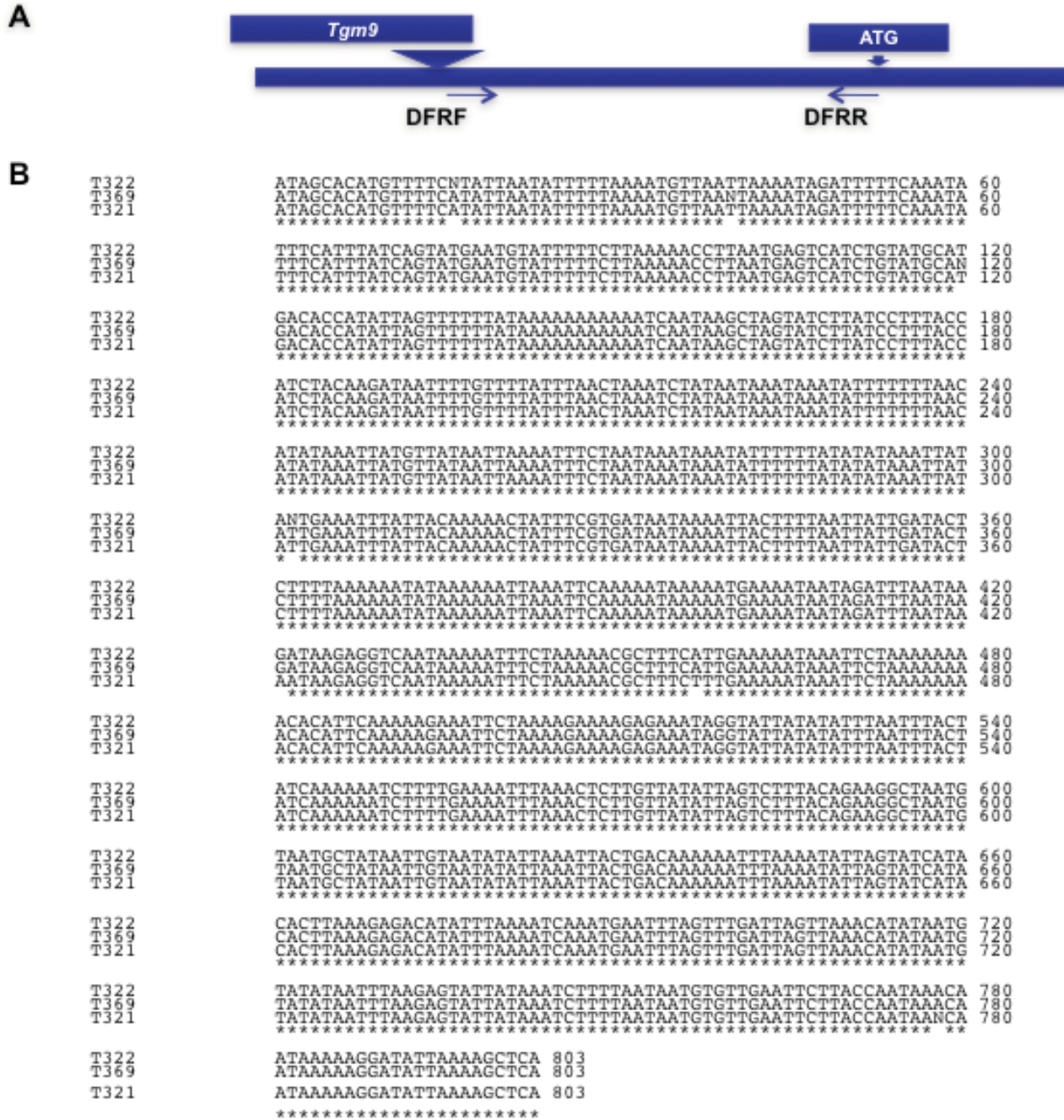


FIGURE S6. —Promoters of *w4-dp* (T321) and *w4-p* (T369) alleles were intact following insertion of *Tgm9*. (A) Schematic representation of the promoter region amplified by PCR is shown. Inverted triangle showed the *Tgm9* insertion site. (B) PCR amplified promoter sequences of parental line T322 (*w4-m*) and two stable mutants *w4-dp* (T321) and *w4-p* (T369) are compared. Primer DFRF, CCTATGCCATGTGAGAATAAAGCAG; Primer DFRR, CCGTATGAAGTGGGTGCTTTTATAG.

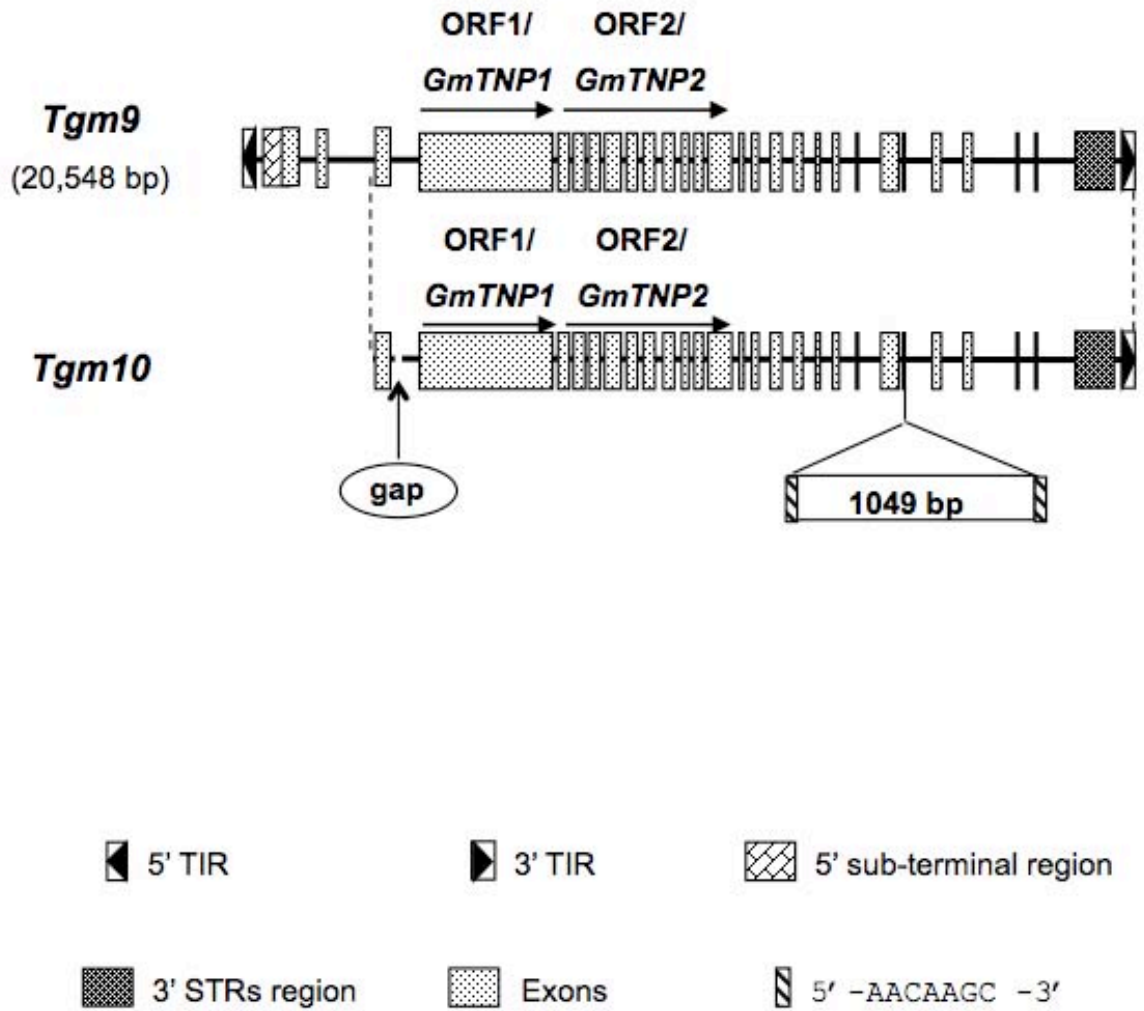


FIGURE S7.—Schematic representation of *Tgm10*. *Tgm10* is located at the 3' end of Scaffold\_57. Except for a ~4100bp deletion in 5' end, a gap in 5'end, and a 1049 bp insertion flanked with a 7 bp direct repeat in exon XXIII, *Tgm10* is 99% identical to *Tgm9* element.

**TABLE S1**  
**Primers used in this study**

Primer <sup>a</sup>	Sequence
ANS1S	5'-ATGCACCTTGGTGAACCATGG-3'
DFR1S	5'-ATCTTGCTGAAGAGGGAAGC-3'
DFR2S	5'-ACAACGAGAGAGAGAAACATG-3'
DFR3S	5'-AGCCTACAATAAACGGATTG-3'
DFR4S	5'-CCCGTTCTTCTATCTTTTTTCTG-3'
Tn3'1S	5'-GGTTGCAGGAGAAAACCGTCTTAGTATGTC-3'
Tn3'2S	5'-CGGTTTTTCGTAACAATCGTC-3'
P1	5'CCCAATCCAAGCTGCGACCTTCAAAG-3'
P3	5'-GGACGAGCCTTTCATCATGGCAGCAC-3'
ANS1R	5'-AATGCTCTTGTACTTGCCGTTG-3'
DFR1R	5'-TGTACATGTCCTCTAAGCTG-3'
DFR2R	5'-TCCCTCTTGAGCAAGATCAG-3'
DFR3R	5'-ATGATATGGTAATGGGACTC-3'
DFR4R	5'-GGACAAAGACAATGCAGGTCCACATCGAAG-3'
Tn3'1R	5'-GACATTCATGTATTCACTAGTACAAATAAAG-3'
P2	5'-CGCCAGAACCCACTTTGTAGCGTGAC-3'
P4	5'-GCTCCCATGTTTGCTGCTCCATACCA-3'
P5	5'-CTGATGCAAGTATGCTCCTAAGTAC-3'

<sup>a</sup> Primers used for sequencing the element are not listed here. ANS1F and ANS1R were designed from a partial coding sequence of an *ANS* gene identified from soybean seed coats (AF325853). DFR1F and DFR1R were designed according to the consensus sequence of three legume *DFR* genes (AF167556 from *G. max*; AF117263 from *Lotus corniculatus*; and AY389346 from *Medicago truncatula*).

**TABLE S2****Probes used in this study**

Probe <sup>a</sup>	Description
<i>ANS</i> partial cDNA	cDNA fragment amplified from purple petals of T322 using ANS1F and ANS1R primers.
<i>F3H</i>	An <i>F3H</i> EST clone (BM093886) provided by Dr. R. C. Shoemaker (Ames, IA)
<i>DFR2</i> partial cDNA	cDNA fragment amplified from petals of T322 using DFR1F and DFR1R primers.
DFR 5'	cDNA fragment amplified from petals of T322 using DFR2S and DFR2R primers.
DFR3'	cDNA fragment amplified from petals of T322 using DFR3S and DFR3R primers.
<i>Tgm9</i> 3' end	PCR fragment amplified using primers TN3'2S and TN3'1R from a lambda clone containing the <i>w4-m</i> allele, isolated from the 1st T322 lambda genomic library.

<sup>a</sup> All the probes were labeled with  $\alpha$ -<sup>32</sup>P-dATP using Primer-it II randomly labeling kit (Stratagene, La Jolla, CA).

**TABLE S3****Polymorphic sequences between *Tgm9* and *Tgmt\****

Nucleotide position in <i>Tgm9</i>	Nucleotide in <i>Tgm9</i>	Nucleotide in <i>Tgmt*</i> <sup>a</sup>
293 (5' STR Exon 1)	T	Y
746 (Intron 1)	C	Y
920 (Intron 2)	C	Y
1645 (Intron 2)	T	C
2043 (Intron 2)	T	C
2192 (Intron 2)	T	A
2146 (Intron 2)	T	C
2923 (Intron 2)	T	-
4969 (Intron 3)	A	G
5129 (Intron 3)	T	C
5272 (Intron 3)	T	-
5715 (Intron 3)	T	C
5851 (Intron 3)	T	-
5863 (Intron 3)	T	-
5999 (Intron 3)	T	C
6651 (Exon 4)	A	G
6819 (Exon 4)	C	Y
8010 (Exon 4)	G	R
8019 (Exon 4)	C	Y
8078 (Exon 4)	C	Y
13359 (Intron 16)	T	C
13524 (Exon 17)	A	G
13616 (Exon 17)	A	G
13904 (Exon 18)	G	A
15519 (Intron 21)	T	C
17169 (Exon25)	A	G

<sup>a</sup> "-" represents nucleotide missing

**TABLE S4****Comparison of 3' terminal inverted repeats among soybean transposable elements**

Transposon	3' TIR <sup>a</sup>	Identity to <i>Tgm9</i>
<i>Tgm9</i>	5'- <b>CACTACTACAAATAAAGCTTTTTAAGTCGG</b> -3'	100%
<i>Tgmt*</i>	5'- <b>CACTACTACAAATAAAGCTTTTTAAGTCGG</b> -3'	100%
<i>Tgm-express1</i>	5'- <b>CACTACTACAAAAGAGGTTTTTTAAGTCGG</b> -3'	87%
<i>Tgm1</i>	5'- <b>CACTATTACAAAAAGTAGTTTTAACATCGG</b> -3'	70%
<i>Tgm6</i>	5'- <b>CACTACTACAAAAGCAGTTTTAACATCGA</b> -3'	70%

<sup>a</sup> Nucleotides in bold are identical to the ones in *Tgm9* 3'TIR.