# Dynamics of mobile element activity in chalcone synthase loci in the common morning glory (*Ipomoea purpurea*)

**Mary L. Durbin*, Amy L. Denton*†, and Michael T. Clegg*‡**

*Department of Botany and Plant Sciences, University of California, Riverside, CA 92521

Contributed by Michael T. Clegg, February 21, 2001

**Mobile element dynamics in seven alleles of the chalcone synthase D locus (CHS-D) of the common morning glory (*Ipomoea purpurea*) are analyzed in the context of synonymous nucleotide sequence distances for CHS-D exons. By using a nucleotide sequence of CHS-D from the sister species *Ipomoea nil* (Japanese morning glory [Johzuka-Hisatomi, Y., Hoshino, A., Mori, T., Habu, Y. & Iida, S. (1999) *Genes Genet. Syst.* 74, 141–147], it is also possible to determine the relative frequency of insertion and loss of elements within the CHS-D locus between these two species. At least four different types of transposable elements exist upstream of the coding region, or within the single intron of the CHS-D locus in *I. purpurea*. There are three distinct families of miniature inverted-repeat transposable elements (MITES), and some recent transpositions of Activator/Dissociation (Ac/Ds)-like elements (Tip100), of some short interspersed repetitive elements (SINEs), and of an insertion sequence (InsIpCHSD) found in the neighborhood of this locus. The data provide no compelling evidence of the transposition of the mites since the separation of *I. nil* and *I. purpurea* roughly 8 million years ago. Finally, it is shown that the number and frequency of mobile elements are highly heterogeneous among different duplicate CHS loci, suggesting that the dynamics observed at CHS-D are locus-specific.**

A major finding of the past quarter century is the diversity of mutational processes that shape the evolution of plant genomes. The relatively regular process of nucleotide substitution, which is predominantly responsible for protein evolution, appears to be supplemented by a variety of recombinational processes that yield gene duplications and rearrangements (1). In addition, a remarkable diversity of mobile elements have been shown to populate plant genomes and to be sources of mutational novelty (2–10). Transposable elements are categorized into two major classes, depending on the mode by which copy number is regulated and on the mechanism of transposition (11). Class I elements transpose via an RNA intermediate and can obtain quite high copy number in a genome. Class II elements move by a ''cut and paste'' mechanism and typically do not occur in high copy number. It is well established that transposon insertions sometimes create novel alleles that correlate with altered levels or altered timing of gene expression (12–14), so these events can have phenotypic effects. It follows from these observations that some subset of insertions may provide a source of adaptive novelty, so transposon dynamics are relevant to the processes of adaptive evolution.

Yet, despite a long history of research on plant mobile elements, relatively little is known about the evolutionary dynamics of transposon behavior within plant species. An exception is the intriguing finding that some Class I transposable elements may have expanded in an episodic fashion over evolutionary time. Thus, recent work suggests that the maize genome has been subject to periodic invasions by retroelements (15). Earlier work had also suggested episodic expansions of the *Ds 1* family of Class II elements in maize and *Tripsacum* (16). In addition, Kalendar *et al.* (17) describe 3-fold variations in copy number of the BARE-1 family of long terminal repeat-retrotransposons within a single population of wild barley in Israel. The barley data reveal a correlation between habitat and retrotransposon number, suggesting that genome evolution may be influenced by local environmental conditions. The causes and genetic consequences of these changes in transposon number is a matter of speculation. Moreover, we know little about the relative differences in dynamic behavior among the various types of mobile elements that appear to reside within plant species. Our goal in this article is to explore this latter question by studying allelic diversity in the chalcone synthase D (CHS-D) region of the common morning glory *Ipomoea purpurea* (L.).

Chalcone synthase controls the first committed step in the flavonoid biosynthetic pathway by condensing three molecules of malonyl CoA with one molecule of 4-coumaroyl CoA to form the 15-carbon naringenin chalcone molecule. The naringenin chalcone product is subsequently modified through the various enzymatic steps of the pathway to produce a wide range of compounds important in floral pigmentation, UV protection, disease defense, and other aspects of phenotype (18). Chalcone synthase is typically represented in plant genomes as a small gene family. Thus, in *Ipomoea*, for example, there are at least six CHS loci. These six copies are grouped into two subfamilies that duplicated and diverged deep in flowering plant evolution (19, 20).

Genetic redundancy affords the opportunity to ask whether transposon insertions are heterogeneous among duplicate gene copies separated by various lengths of evolutionary time. For instance, we can ask whether different duplicate loci harbor distinct types of transposable elements or ask whether the frequency distribution of insertion events is heterogeneous among duplicate loci. An affirmative answer to either of these questions would suggest locus-specific properties that influence insertion dynamics. In addition, the examination of allelic diversity within *I. purpurea* in the context of a closely related outgroup species allows us to infer the temporal history of some insertion events, which provides additional insight into transposon dynamics. In this investigation we use the closely related species *Ipomoea nil* (21) as an outgroup based on the detailed characterizations of insertion elements in the *I. nil* CHS-D locus provided by Shigeru Iida and his colleagues at the National Institute for Basic Biology in Okazaki, Japan (5). Finally, these data allow us to investigate whether insertion dynamics differ among the various types of mobile elements that reside in the CHS-D locus.

---

Abbreviations: CHS, chalcone synthase; MITE, miniature inverted-repeat transposable element; SINEs, short interspersed repetitive elements; UTR, untranslated region.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AF358654–AF358659).

†Present address: Department of Biology, University of Alaska, Fairbanks, AK 99775-6960.

‡To whom reprint requests should be addressed. E-mail: michael.clegg@ucr.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

The samples of CHS-D alleles examined in this article derive from collections of *I. purpurea* from the native range of the plant in Mexico and from the southeastern United States where the plant is an introduced weedy species (20). The common morning glory is distinguished by a number of showy flower color polymorphisms in the United States that have been the subject of extensive population genetic research (22). It is believed that the flower color polymorphisms were selected by native American peoples long before the Spanish invasion of Mexico, and these various color forms have subsequently been propagated as horticultural plants throughout the world.

In this article we show that a remarkable variety of mobile elements reside in the CHS-D region of *I. purpurea.* Our characterization of seven alleles from United States and Mexican populations of *I. purpurea* reveals that every allele is distinct with respect to mobile element pattern. In addition, we show that some classes of elements have exhibited little or no mobility in the time since the separation of the *I. nil* and *I. purpurea* lineages (roughly 8 million years), whereas other classes appear to have been mobilized quite recently. Finally, comparisons among CHS-D, CHS-E, and CHS-A loci reveal substantial heterogeneity among duplicate CHS genes in the presence and distribution of mobile elements.

## Materials and Methods

**PCR.** To amplify the 5′ untranslated region (UTR) and coding sequence of CHS-D, the forward primer was designed at an *Eco*RI site −1091 bp from the ATG start site of the published genomic sequence for CHS-D from *I. purpurea* (GenBank accession no. AB004905): 5′-GAATTCTTATTTAGGTACTTCAATTCATACCATCC-3′. The reverse primer includes 22 bp of the end of exon II and extends 5 bp past the stop codon: 5′-CGGGCTTATGCTGGGACGCTATGGAGG-3′. Amplifications were performed with the Advantage Genomic PCR Kit from CLONTECH. PCR products were cloned by using the TOPO TA Cloning Kit PCR II-TOPO Vector from Invitrogen. Plasmids were purified with Qiagen (Chatsworth, CA) QIAprep Spin Miniprep Kit and sequenced by using the SequiTherm EXCEL II DNA Sequencing Kit-LC from Epicentre Technologies (Madison, WI). The sequencing reactions were analyzed on a Li-Cor dnaSequencer Long ReadIR 4200 (Lincoln, NE). Three independent clones of each PCR product were sequenced one or more times on each strand to provide a minimum of 6-fold coverage for each of the six alleles. On average, 5 kb of sequence was determined for each allele.

**Sequence Analysis.** Sequences were assembled by using SEQUENCHER 3.1 from Gene Codes Corporation (Ann Arbor, MI), and aligned with CLUSTAL V (23). Final alignment was made by visual inspection. The MEGA program (24) was used to calculate genetic distances among sequences. DNASP 3.0 was used to calculate $\theta$ and $\pi$ (25). PAUP 4.0B3A for Macintosh (26) was used to generate a neighbor-joining tree using the Kimura 2-parameter distance estimator (27). Bootstrap support (28) for internal branches was estimated from 1,000 replicates.

## Results

**Nucleotide Sequence Diversity in Exon 1 and 2 of CHS-D.** The gene structure of CHS-D consists of two exons of ≈177 bp and 1088 bp separated by a single intron. A region extending from an *Eco*RI site in the 5′ UTR to the end of exon 2 was amplified and sequenced from six alleles from *I. purpurea* collections. The 5′ UTR differs in size among the alleles ranging from 813 bp to 2002 bp, whereas the introns vary in size from about 3 kb to ≈11 kb depending on the number and type of mobile element insertions. Sequence data from seven CHS-D alleles from *I. purpurea* collected in the southern Mexican states of Oaxaca, Chiapas, and Veracruz (four sequences) and in the southeastern United States (two sequences), together with a published sequence (29), are included in the analyses. Finally, we include a complete CHS-D sequence from a closely related species (*I. nil*) as an outgroup (1).

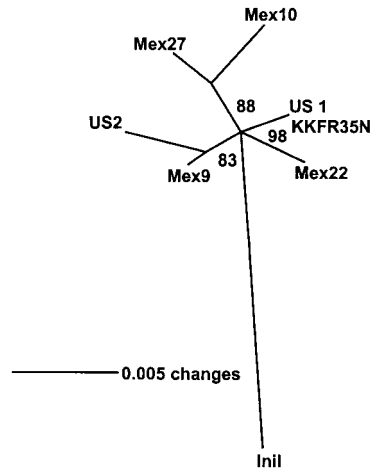


**Fig. 1.** Unrooted neighbor-joining phylogram based on CHS-D coding sequences estimated by using the Kimura 2-parameter model. Numbers below nodes represent bootstrap percentages (1,000 replicates). Sequences for US1 and KKFR35N are identical.

The *I. purpurea* CHS-D exon 1 and 2 data were aligned and analyzed (1104 nucleotides); 23 sites are polymorphic in the sample, eight of which are amino acid replacement changes and 15 of which are synonymous changes. Analysis of the exon sequences indicates a minimum of one recombination event in the history of the sample (30) between sites 27 (exon 1) and 853 (exon 2), but this calculation is biased downwards by the fact that intron sequences were not included in the analysis. Levels of nucleotide sequence diversity ($\theta$ = 0.00850; $\pi$ = 0.00733) are similar to estimates from other plant species (31). Neither Tajima's (32) nor Fu and Li's (33) test statistics are significant in this small sample; although both are negative, suggesting a weak tendency toward an excess of rare alleles. Only Mex9 was heterozygous. This individual contained one unique allele and another allele identical to that found in Mex10. One noteworthy feature of the sample is the complete sequence identity between the US1 and KKFR35N alleles. The identity includes not only the exons, but also the entire intron sequence and 1 kb of 5′ UTR. The KKFR35N allele derived from a horticultural sample sequenced by Habu *et al.* (29). The US1 allele was collected from a natural population in Georgia and described in Epperson and Clegg (34, 35). As noted above, the US populations may have derived from horticultural introductions during the European colonization of this region. The complete sequence identity between these two alleles supports this interpretation because it strongly argues that these two alleles derived from a common ancestor in the relatively recent past. Fig. 1 presents a neighbor-joining tree of the gene genealogy based on the exon data alone.

**Distribution of Mobile Elements in Alleles of CHS-D.** Fig. 2 shows a comparison of seven *I. purpurea* alleles and one *I. nil* allele to illustrate the location and diversity of mobile elements found in the CHS-D region. Both Class I (retrotransposons) and Class II mobile elements (DNA transpons) are found within the CHS-D region. Previously published promoter elements identified in the alleles of CHS-D are also shown in Fig. 2. The insertion sequences in the 5′ UTR occur 5′ to most of the known regulatory elements. Only one regulatory motif (a TACPyAT motif labeled P in Fig. 2) is separated from the other elements

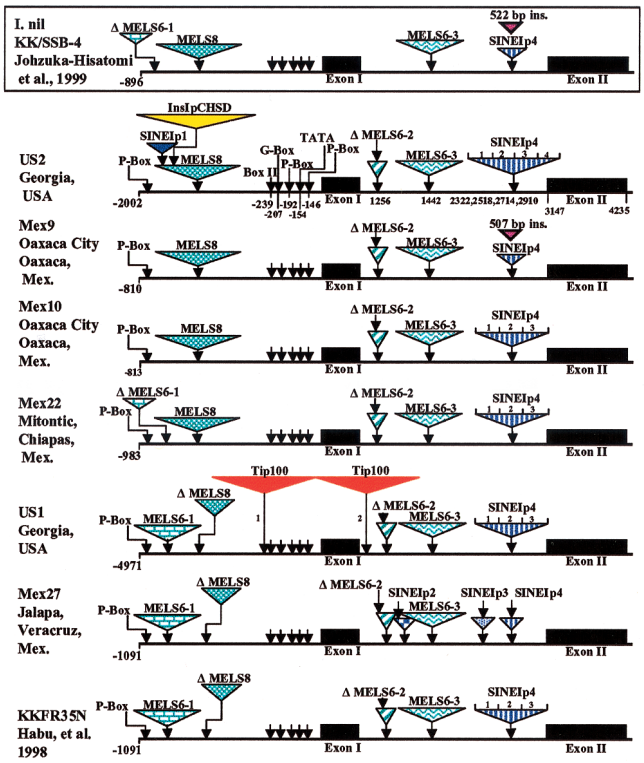


**Fig. 2.** An allele of CHS-D from *I. nil* (boxed) and alleles from individuals of *I. purpurea* collected from populations in the United States and Mexico are diagrammed showing insertion elements and conserved published regulatory elements. The previously published *I. purpurea* CHS-D allele is shown at *Bottom* (KKFR35N).

by the presence of insertion sequences. The region 250 bp 5′ to the start site is highly conserved among all of the alleles of *I. purpurea* and *I. nil*.

We begin with a description of the miniature inverted-repeat transposable elements (MITES) present in both the 5′ flanking region and in the intron (Fig. 2). MITES are relatively short (100–300 bp), have condensed terminal inverted repeats, and often exist as diverse families within the same species (36). MITES are Class II elements. Johzuka-Hisatomi *et al.* (1) reported two families of MITES (MELS3 and MELS6) associated with the CHS-D locus. In this paper we also describe another family of MITES (MELS8). The MELS range in size from 189 bp to about 300 bp and have 10 to 23 bp terminal inverted repeats. There are no discernable target-site duplications flanking the MELS. The MELS have sequence homology to regions in other areas of the genome, such as *I. purpurea* and *I. nil* dihydroflavonol reductase (DFR) noncoding sequence (GenBank accession nos. AB006793 and AB011667) based on sequence similarity identified by using the NCBI BLAST function. MELS6–1 and MELS8 are found at the 5′ end of the 5′ UTR, whereas ΔMELS6–2 (Δ denotes a truncated element) and MELS6–3 are found in the intron very close to one another. The ΔMELS6–2 is in the opposite orientation to the other MELS sequences and is truncated at the same position in all alleles, suggesting that the truncation event occurred in the common ancestor of these individuals.

MELS6–1 is present in five alleles; the one *I. nil* sequence and four of the *I. purpurea* alleles. Of the alleles shown in Fig. 2, only the Mex27, KKFR35N, and US1 alleles have an intact MELS6–1, whereas Mex22 has a MELS6–1 sequence that is truncated at the same position as that of *I. nil*. This suggests that the shared truncation event may have been polymorphic in the common ancestor of *I. nil* and *I. purpurea*. To further examine this possibility, we calculated the pairwise Jukes–Cantor nucleotide substitution distances (K) between alignable portions of the MELS6–1 element. The average estimate (with standard errors in parentheses) is K = 0.0976 (0.018). The average pairwise synonymous distance for exons between the *I. nil* allele and the *I. purpurea* alleles is K = 0.0819 (0.0148), which supports the hypothesis that the MELS6–1 insertion predated the *I. nil*/*I. purpurea* separation. US2, Mex9, and Mex10 do not have a MELS6–1 insertion. If, as the data suggest, the MELS6–1 was present in the common ancestor of *I. purpurea* and *I. nil*, its absence in some alleles (US2, Mex9, and Mex10) may indicate polymorphism for the MELS6–1 element in the common ancestor of these two species. This hypothesis, though plausible, must account for the low probability that a neutral polymorphism would persist for the 8-million-year interval since separation of the lineages. Moreover, reference to Fig. 1 suggests that at least two deletions of MELS6–1 are required to account for the present distribution among *I. purpurea* alleles: also an unlikely event.

The MELS8 MITE has been described by Hoshino *et al.* (49) as Rep. A and Rep. B to designate repetitive sequence. MELS8 seems to be a mobile element because (*i*) there is strong sequence similarity to other *Ipomoea* genes [DFR-A,B,C noncoding regions and a receptor-like protein kinase gene (GenBank accession no. U77888)], and (*ii*) Southern blots indicate a high copy number for this sequence thoughout the *Ipomoea* genome (M.L.D. and M.T.C., unpublished data). The exact 3′ termini of MELS8 are ambiguous because of the age of the original insertion event and the short terminal repeats (10 bp). Further, evidence that MELS8 is a discrete element is that this region is entirely lacking in both *Ipomoea amnicola* and *Ipomoea tricolor* (M.L.D., A.L.D., and M.T.C., unpublished data). The sequence on either side of MELS8 has good sequence alignment with *I. amnicola* and *I. tricolor*. The MELS8 element is about 355 bp and is bounded by 10-bp imperfect terminal inverted repeats. There is no discernable target site duplication. Within the MELS8 is a 13-bp motif that has been identified as part of a promoter for CHS-A in Petunia (Eukaryotic Promoter Database 35055). The average estimate of the Jukes–Cantor distance between MELS8 for *I. nil* and the *I. purpurea* alleles is K = 0.1393 (0.0258), which is higher than the synonymous exon distances and supports the hypothesis that the MELS8 insertion predated the *I. nil*/*I. purpurea* separation.

Inserted within the MELS8 in the 5′ UTR of the US2 allele is a newly described short interspersed repetitive element (SINE), termed SINEIp1 (for SINE *I. purpurea*). SINEs are Class I elements. They do not excise and they are often found in very high copy number (37). SINEIp1 is about 250 bp and contains the split RNA polymerase III promoters and a 16-bp target site duplication characteristic of SINEs (38). US2 is the only allele in which this particular SINE element has been identified (see further discussion below). Within 21 bp of the SINEIp1, and also embedded in MELS8, is a 933-bp insertion termed InsIpCHSD. The InsIpCHSD element is flanked by a 15-bp inverted terminal repeat. The InsIpCHSD element does not have a discernable target site duplication and it has no sequence similarity to any previously published insertion sequences. InsIpCHSD does not contain an ORF. SINEIp1 and InsIpCHSD appear to represent two separate insertions that have occurred since the separation of *I. nil* and *I. purpurea*. It is unclear whether the two events were correlated in time.

In addition to the MITES in the 5′ UTR there are two other types of related MITES in the intron. One is the truncated ΔMELS6–2 and the other is the intact MELS6–3. The ΔMELS6–2 element is absent from the *I. nil* intron, but is present in all *I. purpurea* alleles. There is no discernable footprint in *I. nil* to indicate that the ΔMELS6–2 was present and subsequently

deleted. The average Jukes–Cantor distance between the ΔMELS6–2 sequences is K = 0.0307 (0.0116). Three possible scenarios can account for the absence of ΔMELS6–2 in *I. nil*. One is loss from the *I. nil* lineage subsequent to the *I. nil*/*I. purpurea* separation, the second is ancestral polymorphism in the common ancestor for ΔMELS6–2, and the third is gain by the *I. purpurea* lineage shortly after the separation event so that all descendant *I. purpurea* alleles possess the ΔMELS6–2 insertion.

All of the *I. purpurea* alleles and the *I. nil* allele have the MELS6–3 insertion in the intron. The average distance between *I. nil* and *I. purpurea* MELS6–3 sequences is K = 0.0970 (0.0190). The average estimate of K for the *I. purpurea* alleles with respect to MELS6–3 is 0.0414 (0.0119). These calculations appear to be most consistent with a pattern of common descent, especially in view of the identical spatial location of the element in the *I. nil* and *I. purpurea* introns.

There is considerable variation in the number of SINEs among the alleles (Fig. 2). The SINEs are similar in length (about 250 bp) and all are flanked by target site duplications, but they are substantially diverged in nucleotide sequence. Beginning at the 5′ end of the UTR, SINEIp1 is found only in the US2 allele and is embedded in the MELS8 MITE. All other SINEs are located in the intron. Mex27 has three SINEs in its intron (SINEIp2, SINEIp3, and SINEIp4), whereas the rest of the alleles contain only the SINEIp4. SINEIp3 in the Mex 27 allele is most closely related to SINEIp1, where K = 0.2339 (0.0370). SINEIp2 and SINEIp4 in the Mex27 allele are more diverged from SINEIp1 (K = 0.3390 and K = 0.5293, respectively). The SINEs are least conserved at the 5′ end of the element. If the first 78 bp of the 250-bp element are omitted from the alignment, the distances between SINEIp1 and the rest of the SINEs are K = 0.3256, K = 0.1377, and K = 0.3858, respectively. The RNA polymerase III promoters, characteristic of SINEs, are found in the SINEIp1 and in SINEIp2 sequences, but are not discernible in SINEIp3 and SINEIp4. There is a 60-bp duplication in the Mex27 SINEIp4 that is not found in any of the other SINEIp4s. The average distance between the 60-bp duplicated region in the Mex27 SINEIp4 is K = 0.0341, indicating a relatively recent duplication event. (The sequence alignments for the SINEs are available on request.)

All the alleles contain one or more SINEIp4 insertions at the 3′ end of the intron. The multiple insertions of SINEIp4 occur as tandem repeats with the entire repeat region bounded by a 16-bp imperfect target site duplication. The average pairwise distance between the target site duplication is K = 0.0812 (0.0824). The number of SINEIp4s ranges from one to four. Moreover, the SINEIp4s in the sample of *I. purpurea* are relatively similar (average K within alleles = 0.0612; average K between alleles = 0.0577). The mechanism most likely causing the expansion or contraction of numbers of adjacent SINEs within an allele is some form of unequal recombination. Under this assumption we would expect elements on different alleles to be relatively homogeneous in their genetic distances, owing to recombinational exchange, as the average data indicate. But the calculation of averages also obscures a hierarchical structure. Observed distances between individual SINEIp4 repeats vary from K = 0.0 to K = 0.0889, suggesting that some SINEIp4 repeats are recent duplicates of preexisting elements within the repeat region, as would also be predicted under an unequal recombination hypothesis. Finally, these calculations, together with the fact that *I. nil* also has a SINEIp4 in the identical intron location, suggest that the original insertion event predates the separation of the *I. nil* and *I. purpurea* lineages. (The average Jukes–Cantor distance between the single SINEIp4 repeat in *I. nil* versus the *I. purpurea* repeats is K = 0.1251). In total then we count three different SINEs within the *I. purpurea* sample not shared with the single *I. nil* sequence. It will be important to examine other independently sampled *I. nil* alleles to determine whether these events are confined to *I. purpurea*, but provisionally we assume that these insertions occurred subsequent to the separation of the two lineages.

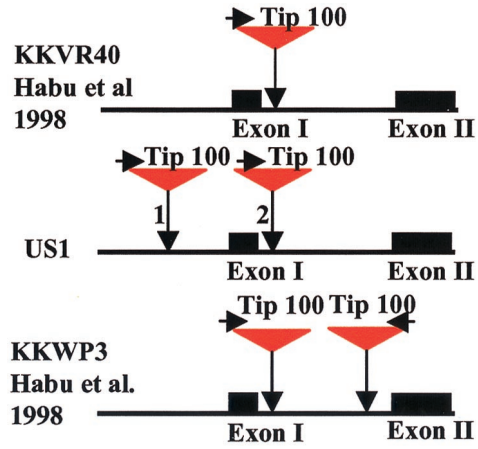


**Fig. 3.** Distribution of Tip100 elements in alleles of CHS-D. Sequence data were not available for the KKVR40 CHS-D and KKWP3 CHS-D alleles and therefore were not included in the data analysis. In addition, sequence data were not available for the Tip100 elements in KKWP3.

Mex9 and *I. nil* both contain an ≈500-bp insertion at the same position in their single SINEIp4 element. There are no inverted repeats or target site duplications associated with the 500-bp insertion. There is also no similarity to any previously published insertion elements. The estimate of pairwise Jukes–Cantor distances between the two 500-bp insertions is K = 0.0903, which is consistent with the divergence estimates between species derived from the exon synonymous site data. This suggests that the insertion predates the separation of *I. nil* and *I. purpurea*; but the fact that the insertion is absent in all other *I. purpurea* alleles also suggests that SINE repeats containing this insertion have tended to be deleted subsequent to the separation of these lineages.

In the 5′ UTR and intron of the US1 allele there is an Activator/Dissociation (Ac/Ds)-like element (Tip100, Figs. 2 and 3). The Ac/Ds elements are Class II elements. The presence of this element blocks the accumulation of transcript for CHS-D, preventing the production of pigment in the outer floral limb (21, 18). Tip100 was described by Habu *et al.* (29) in different alleles of *I. purpurea* (Fig. 3). [Complete sequence data are not available for the KKWP3 and KKVR40 alleles shown in Fig. 3 (29); hence, these alleles were not included in the analyses described above.] The Tip100 element is 3.9 kb with 11 bp terminal inverted repeats and an 8-bp target site duplication. The complete DNA sequence of the Tip100 element in the KKVR40 allele has been published (29), permitting an analysis of Tip100 evolutionary divergence between this allele and US1. One Tip100 is inserted at the identical site in the 5′ end of the intron in all of the alleles shown in Fig. 3. US1 also has a Tip100 in the 5′ UTR that precedes the known regulatory elements. KKWP3 has an additional element inserted at the 3′ end of the intron (sequence data not available for this element). In US1, the Tip100 elements are in the same orientation and the flower exhibits low sectoring rates. In an allele (data not shown) where there is a footprint for the element in the 5′ UTR location (indicating that the Tip100 has excised from this location), the flower exhibits high sectoring rates, suggesting high levels of somatic excision of the remaining element. In the KKWP3 allele, the two elements are in the opposite orientation to one another and the flower is stable white (29), suggesting a negligible level of excision. The two Tip100 elements in the US1 allele differ by a single nucleotide substitution, yielding an average pairwise Jukes–Cantor distance

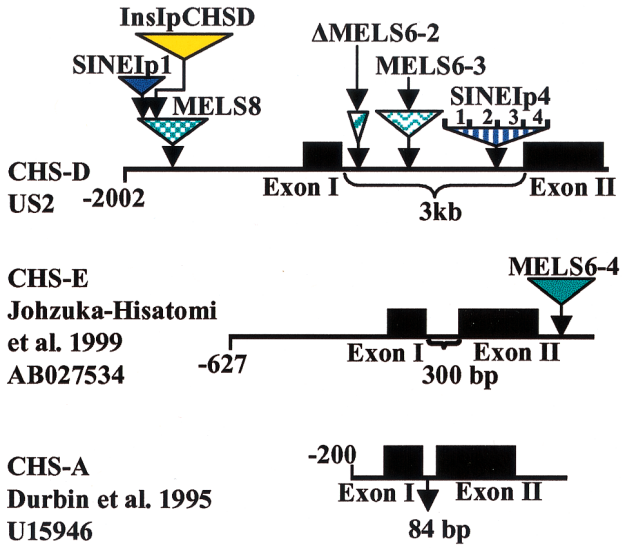


**Fig. 4.** Insertion elements in different *I. purpurea* CHS gene family members.

estimate of K = 0.0003 (0.0003) consistent with a very recent duplication event. The Tip100 element in the 5′ UTR of the US1 allele is identical in nucleotide sequence to the KKVR40 Tip100 element in the 5′ end of the intron (Fig. 2), again suggesting a very recent transposition history. The average nucleotide distance estimate between Tip100 elements is K = 0.0015.

**Frequency of Insertion Events at CHS-D Compared with Other CHS Gene Family Members.** Because the *Ipomoea* genome contains a small family of CHS genes (18, 19), we may ask how the distribution of insertion elements differs among duplicate members of the gene family. Fig. 4 shows insertion elements in three different CHS gene family members in *I. purpurea*. CHS-A is representative of the CHS-A, -B, -C, and PS subfamily (18, 19). Extensive samples of CHS-A from *I. purpurea* in Mexico and the United States failed to reveal any insertion elements in the intron of this gene (39). We also sequenced CHS-A from *I. purpurea* (19) and several other *Ipomoea* species and found their introns to be about 100 bp on average. We sequenced over 1000 bp of the 5′ UTR of CHS-B in *I. purpurea* and did not find the kinds of insertions that are associated with CHS-D. With respect to the CHS-D-E subfamily, only one MITE (MELS6–4) that resides in the 3′ flanking region of CHS-E has been described (1), which is in marked contrast to the high number of elements in the 5′ UTR and intron regions of CHS-D. Correlated with the absence of insertion elements in the introns of CHS-A and -E is the fact that the intron of CHS-D at about 3000 bp is much larger than the ≈300 bp for CHS-E and only 84 bp for CHS-A. These contrasts lead us to infer that the distribution of mobile elements is heterogeneous among CHS genes in the *Ipomoea* genome.

## Discussion

Every allele in the sample differs in insertion pattern, suggesting a remarkable level of insertion and deletion activity within the history of the sample. These observations raise a number of intriguing questions. First, do transposon insertions associated with CHS-D have any discernible effect on plant fitness? Second, do the various kinds of elements inserted within the CHS-D region have different temporal dynamics? And third, is the pattern of transposon insertion heterogeneous across the *Ipomoea* genome? We discuss each of these major questions in turn.

Not surprisingly, the answer to the fitness question appears to depend on the nature of the insertion. The prevalence of the MITES and the SINEIp4 repeats within the sampled alleles would seem to suggest that these insertions have little or no deleterious effect. These elements, which are common to most alleles, seem to have been incorporated into the *Ipomoea* genome before the *I. nil*/*I. purpurea* separation roughly 8 million years ago and it may be that their presence had some beneficial effect on phenotype during or before this transition. It may also be that they were driven to fixation by hitchhiking or because the rate of insertional input was initially very high. Whatever the original forces affecting their dynamics, these elements have diminished in the *I. purpurea* lineage (owing to complete or partial deletions by excision, unequal recombination, or some other mechanism), providing no evidence for a continued advantage (if one ever existed).

What can be deduced about the potential effects of these elements on gene expression? Inspection of the sites of insertion indicates that the great majority of insertion sequences are 5′ to most of the known regulatory elements, which are confined to a highly conserved region of ≈200 bp just proximal to exon 1 (Fig. 2). The exceptions are one TACCAT motif that is separated from the other regulatory elements. The TACPyAT motif is reported to be important in the tissue-specific expression of CHS (40). In addition, a G-Box motif is present in the MELS8 insertion, but this feature is deleted in the Mex27, US1, and KKFR35N alleles. The G-Box has been found to be required for the UV light induction of CHS in *Antirrhinum* and parsley (41). The MELS8 also contains a 13-bp motif that has homology to a region of a promoter for the Petunia CHS-A gene. This motif is 5′ to the other known regulatory elements and it is not known whether its presence has some regulatory effect. MITES found in other species are also known to contain regulatory sequences like putative TATA boxes for transcription and also motifs for polyadenylation (36). More generally, the various elements associated with the CHS-D locus in *I. purpurea* have contributed to the expansion of the 5′ UTR to as much as 2000 bp. This is in contrast to other related *Ipomoea* species that lack these insertion sequences, such as *I. amnicola* and *I. tricolor*, where the comparable 5′ UTR region is only about 300 bp (M.L.D., A.L.D., and M.T.C., unpublished data). Perhaps the extended 5′ UTR also has some subtle effects on gene expression.

Finally, there is direct evidence that insertions of the Tip100 element within the intron confer a phenotypic change (29). Specifically, some Tip100 insertions in the intron disrupt the anthocyanin pathway by blocking CHS-D expression, resulting in an albino floral limb. Moreover, somatic excision of the element restores pigment production, resulting in a sectoring phenotype (18, 29, 35, 42). Other data suggest that there are complex interactions between adjacent Tip100 elements that alter sectoring frequencies. Thus, a low sectoring phenotype appears to be correlated with the presence of two adjacent Tip100 elements, one within the intron and one in the 5′ UTR proximal to exon 1 in the US1 alleles (Fig. 3). Moreover, genetic analyses of sectoring levels by Epperson and Clegg (35, 42) showed that several allelic classes determined rate differences and that these were cis acting and closely linked to the locus responsible for the albino floral limb phenotype. Fig. 3 shows that Tip100 has moved a number of times within the area of the CHS-D locus, indicating that Tip100 is very mobile within the CHS-D locus and surrounding sequences.

The fact that flower color is subject to selection is very well established. A large body of experimental work has shown that white flowers are undervisited by insect pollinators in nature and that this mating discrimination is associated with complex selection dynamics (reviewed in ref. 22). It is also intriguing that the Tip100 insertions are confined to alleles from the United States or horticultural collections. The plants that were introduced into the southeastern United States during the European colonization of this region are believed to have been cultivated forms (20). The selection of these diverse flower color types is

thought to have been the work of pre-Columbian peoples in Mexico who may have selected for Tip100 insertions by selecting the floral phenotypes produced by Tip100 insertions.

The temporal patterns of mobile element insertion seem to differ among element types. A conservative count suggests four complete or partial losses of elements since the *I. purpurea*/*I. nil* separation. These events are accounted for by the MITES. The MITES show no evidence of transposition since the *I. purpurea*/*I. nil* separation. This is consistent with reports that MITEs have not been shown to excise (36). There appear to have been at least seven mobile element gains within the *I. purpurea* lineage that are not shared with the *I. nil* sequence. The gains are accounted for by the InsIpCHSD element, the SINEs, and the Tip100 elements. Based on sequence distances, the movement of Tip100 elements is quite recent. The hierarchy of transposon mobility appears to be Tip100 > InIpCHSD ≈ SINEs > MITES.

A conservative estimate of the rate of gain of new transposons within the CHS-D region can be calculated based on the estimated 8-million-year separation time of the *I. nil* and *I. purpurea* lineages as $8.75 \times 10^{-7}$/yr. Similarly, the estimated rate of loss or partial loss of elements is $5 \times 10^{-7}$/yr. However, these estimates understate the very recent mobility of the Tip100 elements. Assuming a neutral mutation rate of $\approx 7 \times 10^{-9}$ per neutral nucleotide site/year (43), and assuming that the single nucleotide substitution detected in the Tip100 elements are neutral, we calculate that the Tip100 duplications occurred within the last 9,100 years. A minimum of three events are observed in the sample, yielding an approximate rate estimate of $3.3 \times 10^{-4}$ transpositions/year. The actual frequencies could be much higher. Epperson and Clegg (35, 42) estimated excision rates to be $\approx 1 \times 10^{-2}$, based on the direct recovery of stable albino progeny from sectoring maternal parents.

Our final question concerns heterogeneity in mobile element frequency between duplicate copies of CHS genes. It is evident that the CHS-D locus has a much higher frequency of insertion elements within the intron than do either the CHS-A or -E loci. We have the most comprehensive population data with regard to intron sequences, but preliminary indications are that this is also true for the 5′ UTR—although our data on this point are much less complete. Why should these related genes differ in insertion frequency? One speculation concerns level of expression. CHS-D is the most highly expressed of the CHS genes in *I. purpurea* (18): perhaps high levels of expression promote insertion events, as has been suggested elsewhere (44). But genomic data from *Arabidopsis* suggest a negative correlation between gene-rich regions and the accumulation of transposons (45). Moreover, the related species *I. amnicola* and *I. tricolor* show little or no evidence for mobile element insertions in the 5′ UTR of the CHS-D locus; so this speculation does not appear to be consistent with all of the facts. What is the frequency of mobile element insertions in other loci in the flavonoid pathway in *Ipomoea*? Iida and associates (46–48) have studied the dihydroflavonol reductase (DFR) gene family in detail and find considerable evidence for mobile element insertions within these genes in *I. nil* and *I. purpurea*. So CHS-D does not seem to be unique among flavonoid genes in its propensity to collect insertion elements.

We thank Margaret Kidwell, Sue Wessler, Norihiro Okada, Vanessa Ashworth, Michael Cummings, and Peter Morrell for critical reading of the manuscript and helpful suggestions. This work was supported in part by a grant from the Alfred P. Sloan Foundation.



1. Clegg, M. T. (1999) in *Evolutionary Processes and Theory*, ed. Waser, S. P. (Kluwer, Dordrecht, The Netherlands), pp. 55–64.
2. Bennetzen, J. L. (2000) *Plant Mol. Biol.* **42,** 251–269.
3. Clegg, M. T. (1997) *J. Hered.* **88,** 1–7.
4. Fedoroff, N. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 7002–7007.
5. Iida, S., Hoshino, A., Johzuka-Hisatomi, Y., Habu, Y. & Inagaki, Y. (1999) *Ann. N.Y. Acad. Sci.* **870,** 265–274.
6. Kidwell, M. G. & Lisch, D. R. (2000) *Trends Ecol. Evol.* **15,** 95–99.
7. Le, Q. H., Wright, S., Yu, Z. & Bureau, T. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 7376–7381.
8. Lonnig, W. E. & Saedler, H. (1997) *Gene* **205,** 245–253.
9. McClintock, B. (1987) *Stadler Genet. Symp. Ser.* **10,** 25–48.
10. Wendel, J. F. & Wessler, S. R. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 6250–6252.
11. Kunze, R., Saedler, H. & Lönnig, W.-E. (1997) *Adv. Bot. Res.* **27,** 332–470.
12. Che-Hong, C., Oishi, K. K., Kloeckener-Gruissem, B. & Freeling, M. (1987) *Genetics* **116,** 469–477.
13. Kloeckener-Gruissem, B. & Freeling, M. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 1836–1840.
14. Selinger, D. A. & Chandler, V. L. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 15007–15012.
15. Sanmiguel, P. & Bennetzen, J. L. (1998) *Ann. Bot. (London)* **82,** 37–44.
16. Gerlach, W. L., Dennis, E. S., Peacock, W. J. & Clegg, M. T. (1987) *J. Mol. Evol.* **26,** 329–334.
17. Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E. & Schulman, A. H. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 6603–6607. (First Published May 23, 2000; 10.1073/pnas.110587497)
18. Durbin, M. L., McCaig, B. & Clegg, M. T. (2000) *Plant Mol. Biol.* **42,** 79–92.
19. Durbin, M. L., Learn, G. H., Huttley, G. A. & Clegg, M. T. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 3338–3342.
20. Glover, D. E., Durbin, M. L., Huttley, G. & Clegg, M. T. (1996) *Plant Species Biol.* **11,** 41–50.
21. Johzuka-Hisatomi,Y., Hoshino, A., Mori, T., Habu, Y. & Iida, S. (1999) *Genes Genet. Syst.* **74,** 141–147.
22. Clegg, M. T. & Durbin, M. D. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 7016–7023.
23. Kumar, S., Tamura, K. & Nei, M. (1994) *Comput. Appl. Biosci.* **10,** 189–191.
24. Higgins, D. G., Bleasby, A. J. & Fuchs, R. (1992) *Comput. Appl. Biosci.* **8,** 189–191.
25. Rozas, J. & Rozas, R. (1999) *Bioinformatics* **15,** 174–175.
26. Swofford, D. L. (1998) PAUP 4.0, Phylogenetic analysis using parsimony, version 4.0b4a (Sinauer Associates, Sunderland, MA).
27. Kimura, M. (1980) *J. Mol. Evol.* **16,** 111–120.
28. Felsenstein, J. (1985) *Evolution* **39,** 137–195.
29. Habu,Y., Hisatomi, Y. & Iida, S. (1998) *Plant J.* **16,** 371–376.
30. Hudson, R. R. & Kaplan, J. L. (1985) *Genetics* **111,** 147–164.
31. Cummings, M. P. & Clegg, M. T. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 5637–5642.
32. Tajima, F. (1989) *Genetics* **123,** 585–595.
33. Fu, Y.-X. & Li, W.-H. (1993) *Genetics* **133,** 693–709.
34. Epperson, B. K. & Clegg, M. T. (1986) *Am. Nat.* **128,** 840–858.
35. Epperson, B. K. & Clegg, M. T. (1992) *J. Hered.* **83,** 405–409.
36. Wessler, S. R. (1998) *Physiol. Plant.* **103,** 581–586.
37. Shedlock, A. M. & Okada, N. (2000) *BioEssays* **22,** 148–160.
38. Ogiwara, I., Miya, M., Ohshima, K. & Okada, N. (1999) *Mol. Biol. Evol.* **16,** 1238–1250.
39. Huttley, G. A., Durbin, M. L., Glover, D. E. & Clegg, M. T. (1997) *Mol. Ecol.* **6,** 549–558.
40. Vandermeer, I. M., Brouwer, M., Spelt, C. E., Mol., J. N. M. & Stuitje, A. R. (1992) *Plant J.* **2,** 525–535.
41. Martin, C. R. (1993) *Int. Rev. Cytol.* **147,** 233–284.
42. Epperson, B. K. & Clegg, M. T. (1987) *J. Hered.* **78,** 346–352.
43. Gaut, B. S., Morton, B. R., McCaig, B. C. & Clegg, M. T. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 10274–10279.
44. Kumar, A. & Bennetzen, J. L. (1999) *Annu. Rev. Genet.* **33,** 479–532.
45. Kaul, S., Koo, H. L., Jenkins, J., Rizzo, M., Rooney, T., Tallon, L. J., Feldblyum, T., Nierman, W., Benito, M. I., Lin, X. Y., *et al.* (2000) *Nature (London)* **408,** 796–815.
46. Inagaki, Y., Hisatomi, Y. & Iida, S. (1996) *Theor. Appl. Genet.* **92,** 499–504.
47. Takahashi, S., Inagaki, Y., Satoh, H., Hoshino, A. & Iida, S. (1999) *Mol. Gen. Genet.* **261,** 447–451.
48. Inagaki, Y., Johzuka-Hisatomi, Y., Mori, T., Takahashi, S., Hayakawa, Y., Peyachoknagul, S., Ozeki, Y. & Iida, S. (1999) *Gene* **226,** 181–188.
49. Hoshino, A., Johzuka-Hisatomi, Y. & Iida, S. (2001) *Gene* **265,** 1–10.