

SIRE1, an Endogenous Retrovirus Family from *Glycine max*, Is Highly Homogeneous and Evolutionarily Young

Howard M. Laten,* Ericka R. Havecker,† Lisa M. Farmer,* and Daniel F. Voytas†

*Department of Biology, Loyola University Chicago; and †Department of Zoology and Genetics, Iowa State University

SIRE1 is unusual among Ty1-*copia* retrotransposons in that it has an additional open reading frame with structural features similar to retroviral envelope proteins between *pol* and the 3' LTR. Here we report the characterization and comparison of eight different *SIRE1* elements derived from a soybean genomic library, as well as *SIRE1* reverse transcriptases from *Glycine soja*. The DNA sequences of the eight *SIRE1* elements are highly homogeneous and share greater than 95% nucleotide identity. Partial sequences obtained from BAC ends are similarly conserved. Phylogenetic analyses resolve two closely related *SIRE1* lineages, and nucleotide changes within and between *SIRE1* lineages have occurred to preserve function. Both the *gag* and the *env*-like genes are evolving under similar levels of functional constraint. Considerable sequence heterogeneity in the form of short duplications was found within the LTRs and in the region between the envelope-like ORF and the 3' LTR. These duplications are suggestive of slippage by reverse transcriptase during replication. Sequence identity between LTRs of individual insertions suggests that they transposed within the past 70,000 years. Two of 10 *SIRE1* insertions examined about Ty3-*gypsy* retroelements. Since the soybean genome harbors more than 1,000 *SIRE1* insertions, the collective data suggest that *SIRE1* has undergone a very recent and robust amplification in soybean.

Introduction

The genomes of most plants harbor a diverse collection of retroelements, including LINE-like retroposons, long terminal repeat (LTR) retrotransposons, and endogenous retroviruses (Wessler, Bureau, and White 1995; Bennetzen 2000; Peterson-Burch et al. 2000; Vicent, Kalendar, and Schulman 2001). The coding sequences of most characterized plant retroelements are highly degenerate and are cluttered with stop codons, frameshifts, and deletions. Autonomous retrotransposition has only been documented for a handful of plant retroelements (Grandbastien, Spielmann, and Caboche 1989; Hirochika et al. 1996a; Agrawal et al. 2001), yet there is clear evidence of retrotransposon amplification in plants within the last several million years (SanMiguel et al. 1998; Vicent, Kalendar, and Schulman 2001). Thus, despite their general degeneracy, a minority of plant retroelements are functional and actively expanding (Pouteau et al. 1991; Hirochika et al. 1996b; Takeda et al. 1998; Jaaskelainen et al. 1999).

Among the plant LTR retrotransposons, families of both Ty1-*copia* and Ty3-*gypsy* elements have been discovered that encode envelope-like proteins (Turcich et al. 1996; Laten, Majumdar, and Gaucher 1998; Wright and Voytas 1998; Kapitonov and Jurka 1999; Laten 1999; Peterson-Burch et al. 2000; Vicent, Kalendar, and Schulman 2001; Wright and Voytas 2002). This suggests that plant genomes—like those of *Drosophila*, other lower animals, and vertebrates (Boeke and Stoye 1997)—harbor endogenous retroviruses. Of the plant elements with *env*-like genes, *SIRE1*-1 from *Glycine max* is the only element whose coding sequences are not truncated or peppered with nonsense and/or frameshift mutations (Laten, Majumdar, and Gaucher 1998; Laten 1999; Peterson-Burch

et al. 2000). Furthermore, Southern hybridization analyses suggested that the approximately 1,000 *SIRE1* copies in the soybean genome are homogeneous (Laten and Morris 1993; Laten, Majumdar, and Gaucher 1998), implying that *SIRE1* elements may have recently undergone a replication burst. To explore this hypothesis and to understand the evolutionary dynamics of the *SIRE1* family, we characterized the DNA sequences of seven additional independent *SIRE1* insertions, as well as *SIRE1* reverse transcriptases from the ancestral wild soybean, *Glycine soja*.

Materials and Methods

Clones containing *SIRE1* sequences were recovered from a λ genomic library (Stratagene) by plaque hybridization (Sambrook, Fritsch, and Maniatis 1989) using a probe encompassing the integrase (IN) and reverse transcriptase (RT) coding regions and most of the *env*-like gene from *SIRE1*-1 (Laten, Majumdar, and Gaucher 1998). DNAs were isolated from plate lysates (Qiagen) and amplified by standard protocols using recombinant Taq DNA polymerase (Life Technologies). Primer pairs were designed to amplify either the 5' or the 3' end of *SIRE1*-1 to screen for phage clones carrying full-length *SIRE1* elements. The 5' ends were amplified using a LTR forward primer (TGGAAGGTTGTAAACAGTGGC) and a *gag* reverse primer (AGTCGAAAGGGATGTTCCG); 3' ends were amplified using an *env*-like ORF forward primer (ACATTGTCTCGACACAGGG) and a LTR reverse primer (ATATTTTCGGGCAGATG).

For sequencing, phage DNAs were isolated from plate lysates (Qiagen). *SIRE1*-7, *SIRE1*-8, and *SIRE1*-9 DNAs were sequenced directly from recombinant phage at the University of Chicago Cancer Research Center DNA Sequencing Facility, as were selected regions of *SIRE1*-2, *SIRE1*-4, *SIRE1*-13, and *SIRE1*-14. Phage DNAs for *SIRE1*-2, *SIRE1*-4, *SIRE1*-13, and *SIRE1*-14 were amplified using the high fidelity DNA polymerase, Pfx (Invitrogen), with primers based on the sequence of *SIRE1*-1. PCR products were purified (Qiagen) and sequenced at the Iowa State University DNA Sequencing

Key words: endogenous retrovirus, retrotransposon, phylogenetics, Ty1-*copia*, *Glycine max*.

E-mail: hlaten@luc.edu.

Mol. Biol. Evol. 20(8):1222–1230. 2003

DOI: 10.1093/molbev/msg142

Molecular Biology and Evolution, Vol. 20, No. 8,

© Society for Molecular Biology and Evolution 2003; all rights reserved.

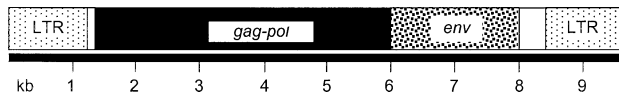


FIG. 1.—Organization of *SIRE1-4*.

Facility. Repeated amplifications using the same templates yielded products with identical DNA sequences. DNAs isolated from *Glycine soja* leaves (Klimyuk et al. 1993) were PCR-amplified under standard conditions using forward and reverse primers based on *SIRE1-1 rt* (GAGGCACTGACTGATGAGTTC and TTCTTTGCAT-*ACTTGCTTTGTGAG*, respectively). PCR products were cloned into TOPO2.1 vectors (Invitrogen), and three clones were sequenced using vector primers at the Iowa State University DNA Sequencing Facility.

DNA sequences were aligned using ClustalW (Higgins, Thompson, and Gibson 1996). The presence of size polymorphisms in the region between the *env*-like ORF and the 3' LTR (bases 8200 to 8700) made alignments difficult, and so the region was manually realigned. Gaps were inserted to maximize alignments of nearly identical blocks of duplicated nucleotides. Phylogenetic and molecular evolutionary analyses were conducted using MEGA version 2.1 (Kumar et al. 2001). DNA p-distances were used for closely related distances ($d < 0.05$) and, where appropriate, gamma distances were calculated using Kimura's two-parameter method (Kimura 1980). Minimum-evolution, neighbor-joining, and maximum-parsimony trees were constructed and were evaluated on the basis of 5,000 bootstrap replicates. To evaluate the synonymous to non-synonymous substitution ratios (d_S/d_N), ORF1 was split into two subregions, one encoding just the structural Gag protein(s) and one encoding PR, IN, and RT (Pol). Since the cleavage site between Gag and Pol is yet unknown, the junction was defined to be 25 codons upstream of the conserved Asp-Ser-Gly presumed to be the protease active site. This position approximates the protease cleavage site for HIV (Pearl and Taylor 1987) as well as for Ty1 (Merkulov et al. 1996) and Ty3 (Kirchner and Sandmeyer 1993). To evaluate the d_S/d_N ratios for the *env*-like ORF, the amino acid immediately following the *pol* termination codon was designated the start codon. Codon-aligned nucleotide sequences were analyzed using SNAP (Nei and Gojobori 1986). The ages of selected elements were deter-

mined by calculating p distances between the two LTRs of individual elements and by using a cruciferous molecular clock to estimate the times of insertion (Haubold and Wiehe 2001). Sequences in GenBank related to *SIRE1* and those flanking *SIRE1* insertions were sought using BlastN, TblastN, and TblastX (Altschul et al. 1997).

Potential TATA promoter elements and transcriptional start sites were predicted using time delay neural network (TDNN) (Reese 2001) and ProScan 1.7 (Prestridge 1995). Transmembrane peptides were identified using TMpred (Hofman and Stoffel 1993) and PHDhtm (Rost et al. 1995). Searches for potential transcription factor binding sites were performed using MatInspector V2.2 based on the TRANSFAC 4.0 database (Quandt et al. 1995) and Signal-Scan (Higo et al. 1999). For clarity and consistency, all nucleotide positions refer to the consensus sequence found with the full alignment (see *Supplementary Material* below).

Results

SIRE1 Sequence Diversity

Because *SIRE1-1* is unique among plant retrovirus-like elements in that its coding information does not appear to contain obvious mutations (Laten, Majumdar, and Gaucher 1998), we conducted a survey of additional elements to assess sequence diversity within the family. Of the seven new elements sequenced from a λ genomic library, two (*SIRE1-4* [Fig. 1] and *SIRE1-8*) were full-length, comprising 9,805 and 9,255 bp, respectively (table 1). *SIRE1-7*, *SIRE1-9*, and *SIRE1-13* are nearly complete copies of 9,072 bp, 9,352 bp, and 8,743 bp, respectively (table 1). However, all three are truncated in the 3' LTR by the cloning vector. *SIRE1-2* is interrupted in *gag* by the cloning vector and *SIRE1-14* is interrupted in *gag* by a repetitive DNA. Only short lengths of *SIRE1-3*, *SIRE1-6*, and *SIRE1-10* were sequenced and are not included in the analysis.

Sequences were aligned in their entirety by ClustalW, and neighbor-joining, minimal-evolution (ME), and maximum-parsimony trees were generated. The ME trees are shown in figure 2. Subtrees were also constructed to independently evaluate the evolutionary histories of the *gag-pol* and the *env*-like genes. A closely related *Lotus japonicus* element from a BAC clone (Sato et al. 2001 [GenBank Accession number AP004500]) was used as the

Table 1
Summary of *SIRE1* Structural Elements and Coding Regions

Element	Length (bp)	LTR (bp)	ORF1 (codons)	ORF2 (codons)	Post ORF2 (bp)	Target Site Duplication
<i>SIRE1-1</i>	9295	1001	1578 ^b	658 ^b	527	AAATT ^d
<i>SIRE1-2</i>	>7798 ^a	902	>1466 ^{a,b,c}	659	523	TGTTG ^d
<i>SIRE1-4</i>	9805	1194	1575	680	636	CTCAA
<i>SIRE1-7</i>	>9072 ^a	1205	1577	683	632	ATTAC ^d
<i>SIRE1-8</i>	9255	999	1577	656	496	CACAT
<i>SIRE1-9</i>	>9352 ^a	1127	1577	681	615	ATTTG ^d
<i>SIRE1-13</i>	>8743 ^a	1116	1578 ^b	686	609	CATGG ^d
<i>SIRE1-14</i>	>7036 ^a	1000	>1164 ^a	668	531	CAAAG ^d

^a Truncated copy (see text).

^b Contains one nonsense mutation.

^c Contains four frameshift mutations.

^d Deduced from flanking DNA at one end only.

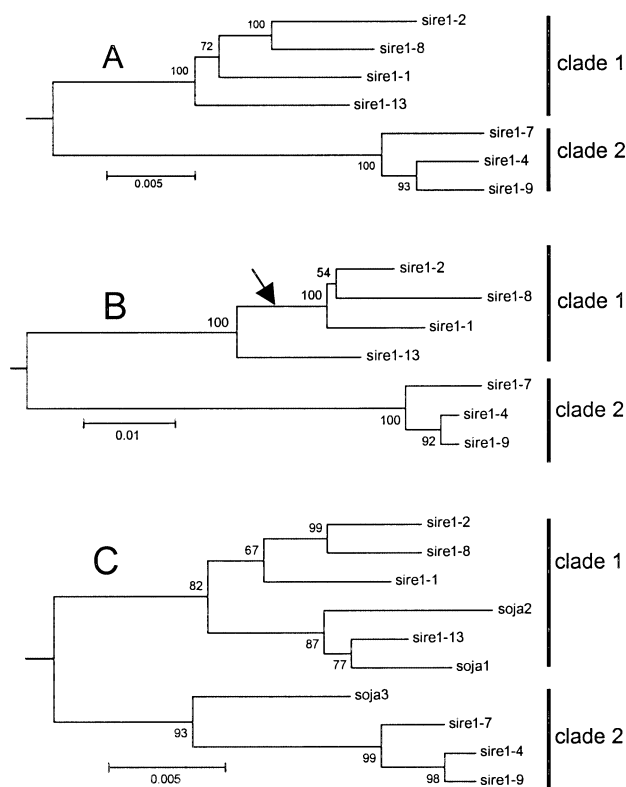


FIG. 2.—*SIRE1* phylogenetic trees based on DNA p-distances, rooted to *L. japonicus* *SIRE1*-related element (see text). Bootstrap percent values at nodes were based on 5,000 bootstrap replicates. (A) ORF1 tree. (B) ORF2 tree. The arrow indicates the internal branch where DNA deletions most probably occurred (see text). (C) *rt* tree.

outgroup. For all sequences, the three methods produced identical, fully bifurcated trees (except as noted below for *SIRE1-14*) that define two *SIRE1* lineages (clades 1 and 2). We amplified by PCR reverse transcriptase domains from *SIRE1* elements from *G. soja*. The ME tree for *rt* is shown in figure 2C. The *G. soja* sequences fall within the two clades and further resolve clade 1 into subfamilies. For *rt*, elements within the clades share greater than 98% nucleotide identity, whereas elements between clades are approximately 96% identical (table 2).

The length variations among these elements for the LTR, the ORF2, and the ORF2-LTR gap define two clearly differentiated groups, one composed of *SIRE1-1*, *SIRE1-2*, and *SIRE1-8* and a second composed of *SIRE1-4*, *SIRE1-7*, *SIRE1-9*, and *SIRE1-13* (tables 1 and 3). The distribution of the specific indels that generate these size classes is consistent with the groupings. The latter group is not monophyletic since *SIRE1-13* is in clade 1 (fig. 2A, B, and C). However, these length disparities can be explained by postulating losses of DNA in the internal branch defining the clade 1 subfamily comprising *SIRE1-1*, *SIRE1-2*, and *SIRE1-8* (see fig. 2B). *SIRE1-13* does not appear to be a recombinant since the informative sites flanking the deletions are consistent with its assignment to clade 1 (data not shown).

SIRE1-14 does appear to be a recombinant element. Analysis of 282 informative sites that unambiguously differentiate clades 1 and 2 define three evolutionary

Table 2
Mean DNA Distances (\pm SE) for *rt* Within and Between Clades

Clade	1 ^a	2 ^a	Lotus ^b
1	0.016 (0.003)	0.036 (0.006)	0.384 (0.038)
2		0.012 (0.003)	0.386 (0.039)
Lotus			NA ^c

^a p-distance values.

^b Kimura two-parameter distance values.

^c NA, not applicable, single sequence.

domains in *SIRE1-14* (fig. 3). Domains I and III define *SIRE1-14* as a sister sequence to *SIRE1-8* in clade 1. The sequence of domain II is identical to the homologous region of *SIRE1-4* (clade 2) except for an insertion of a single triplet.

The high degree of sequence conservation among the sequenced elements was confirmed by analysis of *SIRE1* sequences in GenBank. A BlastN search of the Gene Survey Sequence (GSS) database retrieved 57 additional *SIRE1* elements from sequenced ends of two soybean BAC libraries (Marek et al. 2001). The BAC-end sequences averaged 500 bp in length. Ten overlapping *gag* sequences were 97% identical on average, and the six sequences with similarity to the *env*-like gene shared 93% identity. These values are comparable to the degree of sequence divergence observed for the corresponding regions of the fully sequenced *SIRE1* elements (see below). Forty-eight of the 57 sequences (84%) contained reading frames uninterrupted by stop codons or frameshifts over their entire lengths. In previous work, we estimated that there are approximately 1,000 *SIRE1* copies, which represent 0.5% to 1% of soybean genomic DNA (Laten and Morris 1993). These copy number calculations are consistent with our recovery of 57 *SIRE1* hits from the 6,146 sequences deposited in the GSS database. Hybridizations to arrays of soybean BAC clones also support these estimates (Laten and Meksem, unpublished data).

Another measure of the relative age of the *SIRE1* elements is the divergence between the LTRs of the same element. The LTRs of a single retroelement are theoretically identical at the time of insertion because they are reverse transcribed from the same template sequence. Once integrated, changes in LTR sequences should not be subject to selection, and the frequency should approximate the mutation rate. Of the elements with two complete LTRs, *SIRE1-4* had one base-pair change and *SIRE1-8* had two base-pair changes. The three elements truncated in the 3' LTR: *SIRE1-7*, *SIRE1-9*, and *SIRE1-13*, had no base-pair changes, one base-pair change, and no base-pair changes, respectively. Using an *Arabidopsis* molecular

Table 3
Mean Lengths of Selected *SIRE1* Regions Grouped by Clade (\pm SD)

Elements	Clade	LTR (bp)	ORF2 (bp)	Gap (bp)
1,2,8	1	967 \pm 57	1973 \pm 5	515 \pm 17
4,7,9	2	1175 \pm 42	2044 \pm 6	628 \pm 16
13	1	1116	2058	609

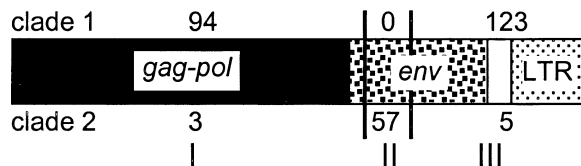


Fig. 3.—Analysis of informative sites in *SIRE1-14*. Distribution of clade-specific sites defining domains I, II, and III. Top numbers are clade 1-specific polymorphisms; bottom numbers are clade 2-specific polymorphisms.

clock for the synonymous evolution rate (Haubold and Wiehe 2001), we estimated that the insertions of *SIRE1-4* and *SIRE1-8* occurred approximately 30,000 and 70,000 years ago, respectively.

The LTRs and Putative *cis*-Acting Sequences

The LTRs range in length from 902 bp to 1,194 bp (table 1). The length polymorphisms among LTRs are due primarily to tandem sequence duplications. The 5' ends of the *SIRE1-4*, *SIRE1-7*, *SIRE1-9*, and *SIRE1-13* LTRs have a common 96-bp duplication separated by 5 bp (fig. 4). The distribution of this duplication replicates that of the length polymorphisms (see table 3). In addition, the LTRs of *SIRE1-4* and *SIRE1-7* have four tandem copies of an imperfect 20-bp repeat beginning at base 726; *SIRE1-13* has three and one quarter copies; *SIRE1-9* has three copies; and *SIRE1-2*, *SIRE1-8*, and *SIRE1-14* have two copies each.

The sequence TATATAA within the LTR was predicted with high confidence to sponsor transcriptional initiation at the adenine at base 630 by both TDNN (Reese 2001) and ProScan (Prestridge 1995) (fig. 4). This location is approximately 300 bp upstream of the 5' end of a previously characterized *SIRE1* cDNA clone (Bi and Laten 1996). The sequence is perfectly conserved among all members except *SIRE1-13*. A conserved sequence candidate for a polyadenylation signal resides upstream of the putative transcriptional start site (base 415 in the 5' LTR). However, a full-length genomic transcript that utilized this site would not contain a repeated region at both the 5' and 3' ends, which is necessary to sponsor strand transfer during reverse transcription. A slightly less favorable candidate for a polyadenylation signal is more appropriately located approximately 200 bp downstream of the proposed transcriptional start site (fig. 4).

The LTRs contain several repeats of variable length that are suggestive of regulatory elements (fig. 4). Although none of these repeats contains motifs resembling *cis*-acting regulatory elements in characterized plant retrotransposons (Grandbastien et al. 1997; Takeda et al. 1999), several contain the sequence AAAG, which forms the core binding site for Dof zinc-finger transcription factors (Yanagisawa and Schmidt 1999). Between bases 418 and 508, this tetranucleotide is present five times (*SIRE1-1*, *SIRE1-2*, *SIRE1-8*, *SIRE1-13*, and *SIRE1-14*) and eight times (*SIRE1-4*, *SIRE1-7*, *SIRE1-9*), respectively. The same sequence is also present at elevated density on the complementary strand (fig. 4). Based on the overall DNA composition of the LTR, AAAG and CTTT would be expected to occur

```
TGTTAGTGCTTAGCACTACTGAGTTTAAAAAGGTTGGCTAAGATTTTGTAAAAACATAAG 60
CACITTAGACAATGAAGGAAAGCTGGAGTTGCTGCACATGATGCCAACGTTATGTCGAAGG 120
AATAAGATCGGGCTGCATAATGCACAGGCGAAGATAAAGTCTCAAGTGAATGAATTGAAGT 180
TGAAGGATCCACCGATGTCGGATACAATGTCCTGACATCCTGCTCGAGAACTACTGGAAGT 240
CTGTACAATGCAAGATAAAGTCAAGTGAAGCATTGAAGCTGCAGGATCCAAGATGTCGG 300
ATACGATGTCCTGACATCTGGCCCGATAAATCTGGACATATAAATCTGTATATCTTTAA 360
CAGATTTATGTCAGTTAGCAAGAGATTAGAAGATCTTCTTTAGGAACGAATAAAGA 420
TCATTAAAGTTTCAATTTCAAGTAGAAGAGTTCGTTCCAGGGATTAAAGATTAAAGATT 480
AAAGATCAAACTAAAGATCAAAAGTTATCTTTTAGTTAGTTCTTTAACTGCAGATTTTTCAGA 540
AGAAGATAGATCTCCTCCAGCATCAAGAACTTGCAGCCAGAAATCGTACACGGCTTATATA 600
ATCATGGAGGCTGCACGAGTTCTGTCCAAAGTCCGGATTGAAGATTAATTTGTGAGTT 660
TTTGGGACTTGAGTCTTTTGTGAGCCACCTTGATGGTACCCTTACATCAAGTGTGGACC 720
TATGTGTGAGAGTTGATCTCTTGTGTAGAGTTGATCTTATTTGTGATGGGTTGATCC 780
CTTTTGTACAGAGTTGATCTCTGATGTGTCTTTGAAATATGTAACACAGAGAGTGTGA 840
GTGAGAGGGAGTGAAGAGGTTCTCATATAGATTTGGGCTTATAGTAGAGATCGCAC 900
GGTAGTGGTTAGTGAGAAGGTTCTAATACAGGGGTTGTAGACCTTGAATCAACTAAT 960
TGAGAGTGGATTCTCCTCCCTGGCTTGGTAGCCCGAGATGAGGTGAGGTTGCCACCGAAC 1020
TGGTAAACAATCTCTTGTGTTATTTACTTGTTTAATCTGTTTCATACGGACACACATA 1080
ACTGCATGTTCTGAAGCATGATGTCGTGACATCTGTACGACATCTGTCGCCCTGGTATCA 1140
GAATTTCA
```

Fig. 4.—LTR from *SIRE1-7* highlighting possible transcriptional elements. Dof-like binding sites are in bold; MYB-like binding sites are in bold italics. Direct repeats are underlined with distinct patterns to differentiate them by sequence. Imperfect tandem repeats of 7 bp and 20 bp, are underlined with short dashes and long dashes, respectively. The putative TATA box is shaded in black, the putative polyA signal is shaded in gray, and the putative RNA start site is indicated by the single A shaded in black.

0.6 and 0.4 times, respectively, in this region. The cluster of AAAG is most dense between 95 and 185 bp upstream of the putative TATA box typical of other retrotransposon regulatory elements (Grandbastien et al. 1997; Takeda et al. 1999).

The tRNA primer binding site (PBS) in *SIRE1* is complementary to soybean tRNA imet (Bi and Laten 1996). Among the insertions sequenced, all clade 1 elements are complementary to 10 bases of the 3' end of the tRNA. Clade 2 elements are complementary to the first 12 bases. Interestingly, the first 10 bases of the PBS (TGGTATCAGA) are repeated just upstream of the 3' end of the LTR in every *SIRE1* member. The polypurine tract (PPT) lies adjacent to the 3' LTR and has the sequence AAAGGGGGAGA. There are no sequence polymorphisms within the PPT or in the 50 bp upstream of this sequence.

gag-pol

A consensus sequence of *SIRE1* elements encodes Gag and Pol on a single open reading frame, which is presumably translated as a single polypeptide. Within Gag-Pol are the invariant amino acid residues and conserved motifs found in most Tyl1-*copia* class retrotransposons (Peterson-Burch and Voytas 2002). These include a zinc finger-like Cys-Cys-His-Cys motif in the presumed nucleocapsid protein (*SIRE1* has two), an Asp-Ser-Gly motif in the catalytic site of protease, His-His-Cys-Cys and Asp-Asp-35-Glu motifs in IN, and several conserved domains within RT.

The *SIRE1 gag-pol* coding region is remarkably conserved, ranging between 95% and 99% identity, with an average of 98%. Some of these nucleotide changes likely compromise *SIRE1* function. *SIRE1-2* has four, single-base frameshift mutations within *gag*, whereas *SIRE1-13* and *SIRE1-1* each have a single nonsense mutation. Despite these obvious mutations, six short indels have occurred that preserve the reading frame. All but one

of these indels are located in the first 1,700 bp of ORF1, within the Gag and PR coding regions. In addition, we calculated the proportion of nucleotide changes that preserved the amino acid sequence (d_S/d_N ratio). For *gag*, defined as the coding region from the presumed start codon to 25 amino acids upstream of the protease active site, the average d_S/d_N ratio among elements was 3.90, denoting selective constraint at most sites. Selection for function of *pol* was considerably stronger, with a d_S/d_N ratio of 7.45.

The *env*-like Gene

The *env*-like gene is in the same reading frame as *gag-pol*, and except for *SIRE1-1*, it is separated from *gag-pol* by a single stop codon. Immediately after the stop codon is a nucleotide motif (CARYTA) known to facilitate stop codon suppression in tobacco mosaic virus (Skuzeski et al. 1991) and several other ssRNA plant viruses (Beier and Grimm 2001). Although there are no examples of Pol-Env fusions in retroelements, constructs carrying the sequence promoted readthrough of the *SIRE1 pol* stop codon in vivo (Havecker and Voytas 2003).

The length polymorphisms in *env* are primarily the result of 11 in-frame indels. All but one of these are confined to the first 550 and last 300 bp of this 2,080-bp ORF. Of the 285 polymorphic nucleotide sites, 25% are located within the first 300 bp of the coding region.

To calculate the d_S/d_N ratio, the nucleotide sequences were codon-aligned, and the ratio was found to average 3.29 between element pairs. Previously, we identified three motifs in the conceptual translation of this ORF analogous to structural elements in retroviral envelope proteins—a transmembrane domain, a fusion peptide, and a coiled-coil domain (Laten, Majumdar, and Gaucher 1998). The putative 19-amino-acid fusion peptide is perfectly conserved among all eight sequenced elements, and the presumed 32-residue coiled-coil has only two polymorphic positions, neither of which alter the heptad repeat pattern (data not shown). The amino terminal transmembrane domain is polymorphic at 16 of 24 residues, yet all variations are predicted to be membrane-spanning peptides with strong confidence (data not shown).

The Interval Between the *env*-like Gene and the 3' LTR

The most variable region in *SIRE1* lies immediately downstream of the *env*-like gene and extends to within 100 bp of the PPT adjacent to the 3' LTR (fig. 5). Variation is primarily in the form of a complex pattern of sequence duplications ranging from simple trinucleotide repeats to imperfect tandem duplications of 100 bp. One shared feature of many of the sequence duplications is the presence of PPT-like sequences. Between bases 8176 and 8845, each *SIRE1* member contains four to six copies of the sequence AGGGGGAG. Another is the presence of short duplications bordering the indels.

Flanking Sequences

We analyzed the DNA adjacent to the *SIRE1* elements in table 1 along with two additional elements

that were partially sequenced, *SIRE1-6* (data not shown) and *SIRE1-10*. Since only *SIRE1-4* and *SIRE1-8* were full-length copies, target site duplications could only be fully evaluated in these two cases. Both are flanked by 5-bp direct repeats: *SIRE1-4* by CTCAA and *SIRE1-8* by CACAT. The 5-bp sequences found adjacent to singular LTRs in the cases of the other seven members are shown in table 1. There does not appear to be a recognizable pattern among these sequences.

SIRE1-1 is adjacent to the *gag-pol* region of a member of the *Ty3-gypsy*-like retroelement, *diaspora* (Yano, Das, Panbehi, Damergis, and Laten, unpublished [GenBank accession number AF095730]). When the sequence adjacent to *SIRE1-10* was translated and used to query GenBank using TblastN and BlastP, the *gag-pol* ORF of *Cinful-1* retroelement sequences from maize (SanMiguel et al. 1996) and *RIRE2* from *Oryza sativa* (Ohtsubo, Kumekawa, and Ohtsubo 1999) were returned, as were several additional closely related sequences from *Lotus japonicus* (Sato et al. 2001). The DNA that abuts the *gag* region of *SIRE1-14* also appears to be a member of a repetitive family. Paralogs of this sequence are present upstream of a coumarate:CoA ligase isoenzyme 3 gene (GenBank accession number AF002257) and a coding region identified as an allergen (GenBank accession number AB013289). The sequence is also represented in 38 BAC-end sequences. The DNA adjacent to the 3' LTR of *SIRE1-14* is not associated with the upstream sequence in either GenBank accession or in the BAC-ends. However, paralogs of this DNA are present in over 20 additional BAC-ends. The 230 bp flanking the 5' end of *SIRE1-4* is 95% identical to a single BAC-end sequence (GenBank accession number AZ221409). None of the other flanking DNAs contained extended ORFs, nor did BlastN or TblastX database searches generate significant hits.

Discussion

Despite the fact that retroelements constitute the majority of chromosomal DNA in many plants, only three retrotransposon families—*Tnt1* and *Tto1* from tobacco and *Tos17* from rice—are known to have members that are transpositionally competent (Grandbastien, Spielmann, and Caboche 1989; Hirochika et al. 1996a; Agrawal et al. 2001). Two lines of evidence support the hypothesis that *SIRE1* may have active members as well. First, the sequence of the originally reported *SIRE1-1* had no obvious frameshifts or stop codons present within its open reading frames. Furthermore, Southern hybridization analysis suggested that most *SIRE1* insertions in the genome were structurally similar to *SIRE1-1* (Laten and Morris 1993; Laten, Majumdar, and Gaucher 1998). Our sequence analysis of additional *SIRE1* insertions more accurately reflect characteristics of the *SIRE1* population and further support the conclusion that this family is highly conserved. In this report we show that the eight sequenced *SIRE1* insertions share greater than 95% nucleotide identity and that 57 additional *SIRE1* sequences from the ends of soybean BAC clones are similarly conserved. Also, sequence divergence between the LTRs of given insertions suggests that some *SIRE1* elements inserted into

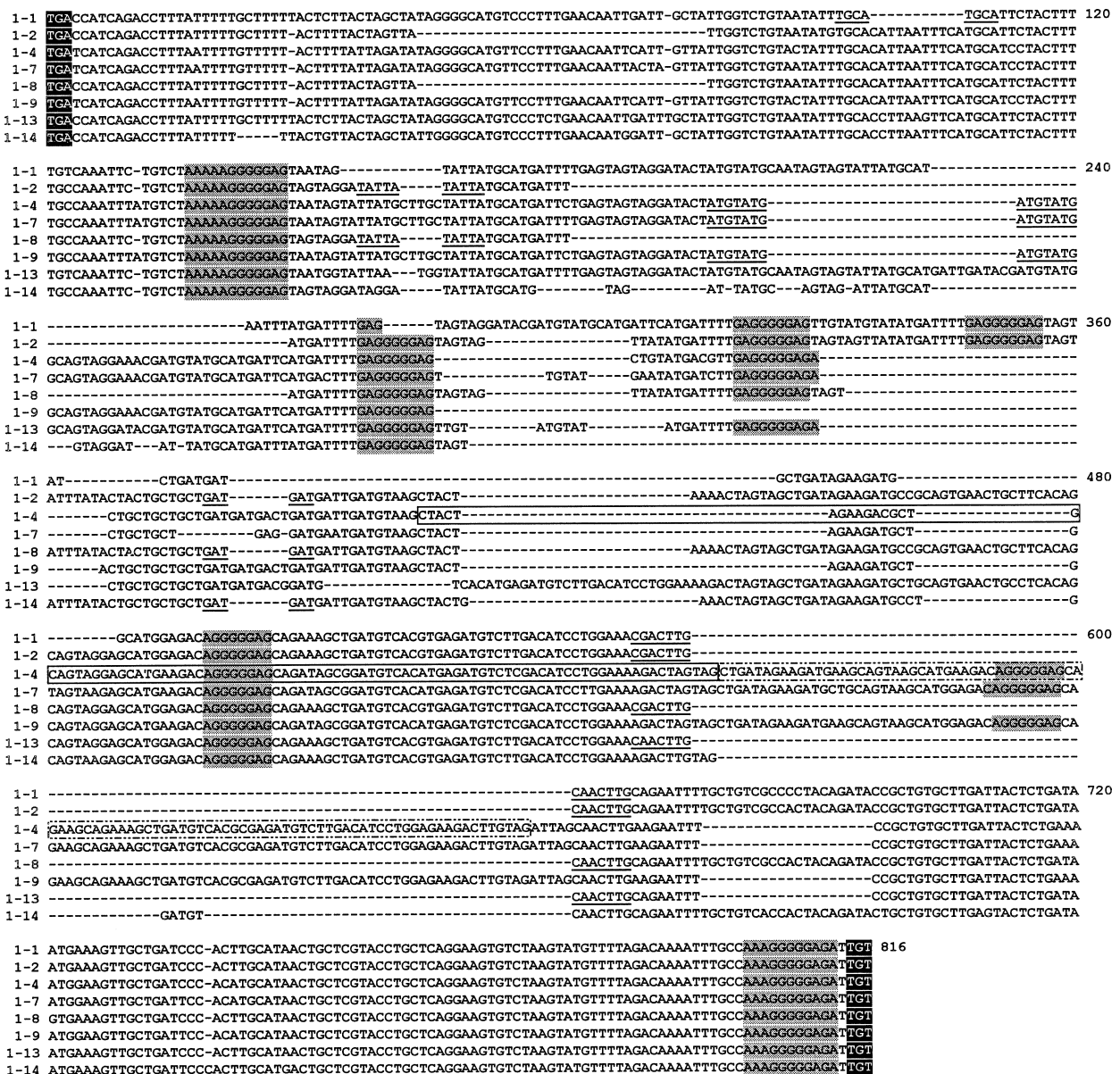


Fig. 5.—Manually modified ClustalW alignment of interval between ORF2 and the 3' LTR. The ORF2 stop codon and the 5' end of the LTR are shaded in black. The PPT and PPT-like tracts are shaded in gray. Short direct repeats that flank indels are underlined. Imperfect long tandem repeat is boxed, with the first repeat in solid lines and the second repeat in dashed lines.

the soybean genome within the past 70,000 years. The level of conservation observed for the *SIRE1* elements is comparable to that reported for the active *Tnt1* family, where up to 4.75% variability between the genomic copies was observed (Casacuberta, Vernhettes, and Grandbastien 1995). The existence of the chimeric member, *SIRE1*-14, suggests that members of both clades have been active simultaneously to generate this recombinant by template switches during reverse transcription.

Of the sequence diversity observed among *SIRE1* family members, most occurs within the noncoding regions, namely the LTRs and the spacer region between the *env*-like ORF and the 3' LTR. Particularly evident are tandem sequence duplications in the 5' portion of the LTR that result in length polymorphisms ranging from 902 to

1,205 bp. In addition, the shorter duplications contained multiple candidate-binding sites for the Dof zinc finger transcription factor just upstream of the putative promoter. Dof proteins regulate a broad spectrum of target genes in both monocots and dicots, including those that are auxin regulated (Kisu et al. 1997; Baumann et al. 1999), light responsive (Yanagisawa and Sheen 1998), and stress induced (Zhang et al. 1995). Stress conditions and defense elicitors are known to induce *Tnt1*, *Tto1*, and *Tos17* (Hirochika et al. 1996b; Grandbastien et al. 1997; Takeda et al. 1998). Repetition of putative, *cis*-acting sequence motifs in LTRs has been noted in four actively transcribed elements: *BARE1*, *Tos17*, *Tnt1*, and *Tto1* (Hirochika et al. 1996b; Suoniemi, Narvanto, and Schulman 1996; Grandbastien et al. 1997; Takeda et al. 1999). In the cases of *Tnt1* and

Tto1, the repeated motifs have been shown experimentally to sponsor inducible element expression (Grandbastien et al. 1997; Takeda et al. 1999), and a MYB-related transcription factor was shown to interact with and regulate *Tto1* at these motifs (Sugimoto, Takeda, and Hirochika 2000). In barley, a MYB transcription factor interacts with the Dof transcription factor, BPBF, to regulate endosperm-specific genes (Diaz et al. 2002). Interestingly, the *SIRE1* LTRs contain two potential MYB-binding sites just upstream of the AAAG-dense region (fig. 4). As of yet, there is no evidence that *SIRE1* RNA is induced by these types of stimuli.

The region between the *env*-like ORF and the 3'LTR varies in length from 496 to 636 bp. The sequence duplications in this region are unusual but not unprecedented among retroelements. The *Grandel* family from maize contains two arrays of tandem repeats between *pol* and the 3' LTR (Martinez-Izquierdo, Garcia-Martinez, and Vicent 1997), and numerous PPT-like sequences characterize the large noncoding region following the *env*-like ORF of the Arabidopsis *Athila* elements (Wright and Voytas 2002). The best explanation for the gain and loss of these repeats is replication slippage (Viguera, Canceill, and Ehrlich 2001). Since strand transfer is a requisite component of retrovirus and retrotransposon replication, some replication slippage by RT at internal regions is quite plausible. Reinitiation at nearby similar or duplicated sequences upstream or downstream could be expected, generating the kind of duplications and subsequent deletions that pervade retroviral genomes (Temin 1993). The presence of tandem triplet repeats and direct repeats of 4 to 7 bp flanking several of the gaps (fig. 5) is consistent with this explanation. In fact, long direct repeats in retroviral DNAs are deleted at high frequency (Rhode, Emerman, and Temin 1987).

With the exception of *SIRE1-2*, sequence variation within *gag-pol* and the *env*-like gene seems to preserve coding information, as all duplications and deletions maintain the open reading frame. For example, the variable N-termini of the *env*-like ORFs among the eight elements contain five different indels, all of which preserve the reading frame. In addition, the high ratio of synonymous to nonsynonymous changes among the *SIRE1* genes further indicates that the elements are evolving under purifying selection. In our study, the d_S/d_N ratio for *pol* averages 7.45, whereas the ratios for *gag* and the *env*-like gene average 3.90 and 3.29, respectively. Interestingly, in the predicted ENV-like protein, amino acid substitutions still preserved structural features of the protein. For example, the predicted N-terminal transmembrane domain was preserved in all *SIRE1* proteins analyzed, despite the relatively high number of nonsynonymous substitutions in this region. The functional constraint placed on the *SIRE1 env*-like gene contrasts with what has been found in mammalian retroviral envelope genes, where adaptive selection results in high levels of variation to avoid the immune response (Nielsen and Yang 1998; Yamaguchi-Kabata and Gojobori 2000).

The flanking DNAs of 10 *SIRE1* insertions were sequenced and two belong to identified plant members of the Ty3-*gypsy* family. Of the remaining eight, one is

flanked on either side by members of two different repetitive families, and one is an apparent paralog of a single BAC-end sequence. The identities of the rest are unknown. These results are suggestive of clustering and/or nesting of some high-copy-number retroelements in *G. max*, similar to what has been reported for other plant genomes (Bennetzen 2000).

Glycine max has been under cultivation for approximately 10,000 years (Hymowitz and Newell 1981) and was derived from wild ancestral *Glycine soja*. Both species contain *SIRE1* elements, indicating that this family was present long before soybean domestication. However, the high degree of similarity among *SIRE1* sequences suggests that this family may still be proliferating in the soybean genome.

The presence of *env*-like ORFs in *SIRE1* and some Ty3-*gypsy* retroelements has raised speculation that these elements may be retroviruses. The functional role, if any, of an envelope protein for viral propagation in a plant host is unknown, and cell walls preclude membrane fusion as a suitable invasive strategy. However, the presence of *env* genes in plant viruses is not unusual. All enveloped plant viruses utilize invertebrate vectors in which the glycosylated envelope proteins sponsor host cell recognition and membrane fusion (VandenHeuvel, Franz, and VanderWilk 2002). ENV has been shown to be dispensable in the plant host. When tospoviruses, plant members of the Bunyaviridae, are maintained solely by mechanical inoculation of host plants, morphological isolates that lack functional envelope proteins can be recovered with point and frameshift mutations in the glycoprotein gene (Goldbach and Peters 1996). These isolates are active in the plant host but fail to reinfect the native thrips host (Goldbach and Peters 1996; Nagata et al. 2000).

The presence of a conserved ORF in a multicopy element family that can be identified across diverse host plant taxa constitutes strong evidence that it has been and may continue to be selectively maintained. Although the conceptual polypeptides from the unusual ORFs of *SIRE1*, *Athila4*, and *Bagy-2* have been designated as *env*-like (Laten, Majumdar, and Gaucher 1998; Vicent, Kalendar, and Schulman 2001; Wright and Voytas 2002), they may embody heretofore unknown functions related to maintenance in their plant hosts. Whatever the role of the *env*-like gene, the high levels of sequence conservation observed among *SIRE1* elements offers promise that these elements can be used to understand the function of this additional coding sequence.

Supplementary Material

All sequences have been deposited into the GenBank database as accession numbers AF053008 (*SIRE1-1*), AY205606 (*SIRE1-2*), AY205607 (*SIRE1-3*), AY205608 (*SIRE1-4*), AY205609 (*SIRE1-7*), AY205610 (*SIRE1-8*), AY205611 (*SIRE1-9*), AY205612 (*SIRE1-13*), AY205613 (*SIRE1-14*), AY212109 (*SIRE1-10*), and AY212110 (*SIRE1-10*). The full *SIRE1* sequence alignment that is summarized in table 1 can be found online at the journal's Web site and at http://www.luc.edu/faculty/hlaten/mbe_data.

Acknowledgments

The authors acknowledge the efforts of E. Gaucher, A. Das, R. Reisner, J. Damergis, B. Panbehi, J. Tziolas, H. Davakos, P. Oh, and T. Fruscione. We thank William Buikema for sequencing expertise and Laura Frederick Marek for helpful suggestions. This work was funded by grants from Loyola University Chicago and in part by Phytodyne, Inc.

Literature Cited

- Agrawal, G. K., M. Yamazaki, M. Kobayashi, R. Hirochika, A. Miyao, and H. Hirochika. 2001. Screening of the rice viviparous mutants generated by endogenous retrotransposon Tos17 insertion: tagging of a zeaxanthin epoxidase gene and a novel OsTATC gene. *Plant Physiol.* **125**:1248–1257.
- Altschul, S. F., T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped Blast and Psi-Blast—a new generation of protein database search programs. *Nucleic Acids Res.* **25**:3389–3402.
- Baumann, K., A. De Paolis, P. Costantino, and G. Gualberti. 1999. The DNA binding site of the Dof protein NtBBF1 is essential for tissue-specific and auxin-regulated expression of the rolB oncogene in plants. *Plant Cell* **11**:323–333.
- Beier, H., and M. Grimm. 2001. Misreading of termination codons in eukaryotes by natural nonsense suppressor tRNAs. *Nucleic Acids Res.* **29**:4767–4782.
- Bennetzen, J. L. 2000. Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* **42**:251–269.
- Bi, Y. A., and H. M. Laten. 1996. Sequence analysis of a cDNA containing the gag and prot regions of the soybean retrovirus-like element, *SIRE-I*. *Plant Mol. Biol.* **30**:1315–1319.
- Boeke, J. D., and J. P. Stoye. 1997. Retrotransposons, endogenous retroviruses, and the evolution of retroelements. Pp. 343–435 in J. M. Coffin, S. H. Hughes and H. E. Varmus, eds. *Retroviruses*. Cold Spring Harbor Laboratory Press, Plainview, New York.
- Casacuberta, J. M., S. Vernhettes, and M. A. Grandbastien. 1995. Sequence variability within the tobacco retrotransposon Tnt1 population. *EMBO J.* **14**:2670–2678.
- Diaz, I., J. Vicente-Carbajosa, Z. Abraham, M. Martinez, I. Isabel-La Moneda, and P. Carbonero. 2002. The GAMYB protein from barley interacts with the Dof transcription factor BPBF and activates endosperm-specific genes during seed development. *Plant J.* **29**:453–464.
- Goldbach, R., and D. Peters. 1996. Molecular and biological aspects of tospoviruses. Pp. 129–157 in R. M. Elliot, ed. *The Bunyaviridae*. Plenum Press, New York.
- Grandbastien, M. A., H. Lucas, J. B. Morel, C. Mhiri, S. Vernhettes, and J. M. Casacuberta. 1997. The expression of the tobacco Tnt1 retrotransposon is linked to plant defense responses. *Genetica* **100**:241–252.
- Grandbastien, M. A., A. Spielmann, and M. Caboche. 1989. Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. *Nature*. **337**:376–380.
- Haubold, B., and T. Wiehe. 2001. Statistics of divergence times. *Mol. Biol. Evol.* **18**:1157–1160.
- Havecker, E. R., and D. F. Voytas. 2003. The soybean retroelement *SIREI* uses stop codon suppression to express its envelope-like protein. *EMBO Rep.* **4**:274–277.
- Higgins, D. G., J. D. Thompson, and T. J. Gibson. 1996. Using Clustal for multiple sequence alignments. *Meth. Enzymol.* **266**:383–402.
- Higo, K., Y. Ugawa, M. Iwamoto, and T. Korenaga. 1999. Plant *cis*-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.* **27**:297–300.
- Hirochika, H., H. Otsuki, M. Yoshikawa, Y. Otsuki, K. Sugimoto, and S. Takeda. 1996a. Autonomous transposition of the tobacco retrotransposon Tto1 in rice. *Plant Cell* **8**:725–734.
- Hirochika, H., K. Sugimoto, Y. Otsuki, H. Tsugawa, and M. Kanda. 1996b. Retrotransposons of rice involved in mutations induced by tissue culture. *Proc. Natl. Acad. Sci. USA* **93**:7783–7788.
- Hofman, K., and W. Stoffel. 1993. TMbase—A database of membrane spanning protein segments. *Biol. Chem. Hoppe Seyler* **374**:166.
- Hymowitz, T., and C. A. Newell. 1981. Taxonomy of the genus *Glycine*: domestication and uses of soybeans. *Econ. Bot.* **35**:272–288.
- Jaaskelainen, M., A. H. Mykkanen, T. Arna, C. M. Vicient, A. Suoniemi, R. Kalendar, H. Savilahti, and A. H. Schulman. 1999. Retrotransposon *BARE-I*: expression of encoded proteins and formation of virus-like particles in barley cells. *Plant J.* **20**:413–422.
- Kapitonov, V. V., and J. Jurka. 1999. Molecular paleontology of transposable elements from *Arabidopsis thaliana*. *Genetica* **107**:27–37.
- Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**:111–120.
- Kirchner, J., and S. Sandmeyer. 1993. Proteolytic processing of Ty3 proteins is required for transposition. *J. Virology* **67**:19–28.
- Kisu, Y., Y. Harada, M. Goto, and M. Esaka. 1997. Cloning of the pumpkin ascorbate oxidase gene and analysis of a *cis*-acting region involved in induction by auxin. *Plant Cell Physiol.* **38**:631–637.
- Klimyuk, V. I., B. J. Carroll, C. M. Thomas, and J. D. Jones. 1993. Alkali treatment for rapid preparation of plant material for reliable PCR analysis. *Plant J.* **3**:493–494.
- Kumar, S., K. Tamura, I. B. Jakobsen, and M. Nei. 2001. MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17**:1244–1245.
- Laten, H. M. 1999. Phylogenetic evidence for Ty1-*copia*-like endogenous retroviruses in plant genomes. *Genetica* **107**:87–93.
- Laten, H. M., A. Majumdar, and E. A. Gaucher. 1998. *SIRE-I*, a *copia*/Ty1-like retroelement from soybean, encodes a retroviral envelope-like protein. *Proc. Natl. Acad. Sci. USA* **95**:6897–6902.
- Laten, H. M., and R. O. Morris. 1993. *SIRE-I*, a long interspersed repetitive DNA element from soybean with weak sequence similarity to retrotransposons: initial characterization and partial sequence. *Gene* **134**:153–159.
- Marek, L. F., J. Mudge, L. Darnielle et al. (19 co-authors). 2001. Soybean genomic survey: BAC-end sequences near RFLP and SSR markers. *Genome* **44**:572–581.
- Martinez-Izquierdo, J. A., J. Garcia-Martinez, and C. M. Vicient. 1997. What makes *Grandel* retrotransposon different? *Genetica* **100**:15–28.
- Merkulov, G. V., K. M. Swiderek, C. B. Brachmann, and J. D. Boeke. 1996. A critical proteolytic cleavage site near the C terminus of the yeast retrotransposon Ty1 Gag protein. *J. Virol.* **70**:5548–5556.
- Nagata, T., A. K. Inoue-Nagata, M. Prins, R. Goldbach, and D. Peters. 2000. Impeded thrips transmission of defective tomato spotted with virus isolates. *Phytopathology* **90**:454–459.
- Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**:418–426.
- Nielsen, R., and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**:929–936.

- Ohtsubo, H., N. Kume-kawa, and E. Ohtsubo. 1999. *RIRE2*, a novel gypsy-type retrotransposon from rice. *Genes Genet. Syst.* **74**:83–91.
- Pearl, L. H., and W. R. Taylor. 1987. Sequence specificity of retroviral proteases. *Nature*. **328**:482.
- Peterson-Burch, B. D., and D. F. Voytas. 2002. Genes of the Pseudoviridae (Ty1/copia retrotransposons). *Mol. Biol. Evol.* **19**:1832–1845.
- Peterson-Burch, B. D., D. A. Wright, H. M. Laten, and D. F. Voytas. 2000. Retroviruses in plants? *Trends Genet.* **16**:151–152.
- Pouteau, S., E. Huttner, M. A. Grandbastien, and M. Caboche. 1991. Specific expression of the tobacco Tnt1 retrotransposon in protoplasts. *EMBO J.* **10**:1911–1918.
- Prestridge, D. S. 1995. Predicting pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.* **249**:923–932.
- Quandt, K., K. Frech, H. Karas, E. Wingender, and T. Werner. 1995. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.* **23**:4878–4884.
- Reese, M.G. 2001. Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. *Comp. Chem.* **26**:51–56.
- Rhode, B. W., M. Emerman, and H. M. Temin. 1987. Instability of large direct repeats in retrovirus vectors. *J. Virol.* **61**:925–927.
- Rost, B., R. Casadio, P. Fariselli, and C. Sander. 1995. Transmembrane helices predicted at 95-percent accuracy. *Protein Sci.* **4**:521–533.
- Sambrook, J., E. F. Fritsch, and T. Maniatis. 1989. *Molecular cloning*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- SanMiguel, P., B. S. Gaut, A. Tikhonov, Y. Nakajima, and J. L. Bennetzen. 1998. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**:43–45.
- SanMiguel, P., A. Tikhonov, Y. K. Jin et al. (11 co-authors). 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274**:765–768.
- Sato, S., T. Kaneko, Y. Nakamura, E. Asamizu, T. Kato, and S. Tabata. 2001. Structural analysis of a *Lotus japonicus* genome. I. Sequence features and mapping of fifty-six TAC clones which cover the 5.4 Mb regions of the genome. *DNA Res.* **8**:311–318.
- Skuzeski, J. M., L. M. Nichols, R. F. Gesteland, and J. F. Atkins. 1991. The signal for a leaky UAG stop codon in several plant viruses includes the two downstream codons. *J. Mol. Biol.* **218**:365–373.
- Sugimoto, K., S. Takeda, and H. Hirochika. 2000. MYB-related transcription factor NtMYB2 induced by wounding and elicitors is a regulator of the tobacco retrotransposon Tto1 and defense-related genes. *Plant Cell* **12**:2511–2527.
- Suoniemi, A., A. Narvanto, and A. H. Schulman. 1996. The *BARE-1* retrotransposon is transcribed in barley from an LTR promoter active in transient assays. *Plant Mol. Biol.* **31**:295–306.
- Takeda, S., K. Sugimoto, H. Otsuki, and H. Hirochika. 1998. Transcriptional activation of the tobacco retrotransposon Tto1 by wounding and methyl jasmonate. *Plant Mol. Biol.* **36**:365–376.
- . 1999. A 13-bp *cis*-regulatory element in the LTR promoter of the tobacco retrotransposon Tto1 is involved in responsiveness to tissue culture, wounding, methyl jasmonate and fungal elicitors. *Plant J.* **18**:383–393.
- Temin, H. M. 1993. Retrovirus variation and reverse transcription: abnormal strand transfers result in retrovirus genetic variation. *Proc. Natl. Acad. Sci. USA* **90**:6900–6903.
- Turcich, M. P., A. Bokhaririza, D. A. Hamilton, C. P. He, W. Messier, C. B. Stewart, and J. P. Mascarenhas. 1996. Prem-2, a copia-type retroelement in maize is expressed preferentially in early microspores. *Sex. Plant Reprod.* **9**:65–74.
- Vandenheuvel, J. F. J. M., A. W. E. Franz, and F. Vanderwilck. 2002. Molecular basis of virus transmission. Pp. 183–210 in C. L. Mandahar, ed. *Molecular biology of plant viruses*. Kluwer, Boston.
- Vicient, C. M., R. Kalendar, and A. H. Schulman. 2001. Envelope-class retrovirus-like elements are widespread, transcribed and spliced, and insertionally polymorphic in plants. *Genome Res.* **11**:2041–2049.
- Viguera, E., D. Canceill, and S. D. Ehrlich. 2001. Replication slippage involves DNA polymerase pausing and dissociation. *EMBO J.* **20**:2587–2595.
- Wessler, S. R., T. E. Bureau, and S. E. White. 1995. LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Curr. Opin. Genet. Dev.* **5**:814–821.
- Wright, D. A., and D. F. Voytas. 1998. Potential retroviruses in plants: Tat1 is related to a group of *Arabidopsis thaliana* Ty3/gypsy retrotransposons that encode envelope-like proteins. *Genetics* **149**:703–715.
- . 2002. Athila4 of *Arabidopsis* and Calypso of soybean define a lineage of endogenous plant retroviruses. *Genome Res.* **12**:122–131.
- Yamaguchi-Kabata, Y., and T. Gojobori. 2000. Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. *J. Virol.* **74**:4335–4350.
- Yanagisawa, S., and R. J. Schmidt. 1999. Diversity and similarity among recognition sequences of Dof transcription factors. *Plant J.* **17**:209–214.
- Yanagisawa, S., and J. Sheen. 1998. Involvement of maize Dof zinc finger proteins in tissue-specific and light-regulated gene expression. *Plant Cell* **10**:75–89.
- Zhang, B., W. Chen, R. C. Foley, M. Buttner, and K. B. Singh. 1995. Interactions between distinct types of DNA binding proteins enhance binding to ocs element promoter sequences. *Plant Cell* **7**:2241–2252.

Pierre Capy, Associate Editor

Accepted March 31, 2003