Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 6897–6902, June 1998
Genetics

SIRE-1, a *copia*/*Ty1*-like retroelement from soybean, encodes a retroviral envelope-like protein

HOWARD M. LATEN*, ARPITA MAJUMDAR, AND ERIC A. GAUCHER

Biology Department, Loyola University Chicago, 6525 North Sheridan Road, Chicago, IL 60626

Edted by Margaret G. Kidwell, University of Arizona, Tucson, AZ, and approved March 23, 1998 (received for review October 16, 1997)

ABSTRACT **The soybean genome hosts a family of several hundred, relatively homogeneous copies of a large, *copia*/*Ty1*-like retroelement designated *SIRE-1*. A copy of this element has been recovered from a *Glycine max* genomic library. DNA sequence analysis of two *SIRE-1* subclones revealed that *SIRE-1* contains a long, uninterrupted, ORF between the 3′ end of the *pol* ORF and the 3′ long terminal repeat (LTR), a region that harbors the *env* gene in retroviral genomes. Conceptual translation of this second ORF produces a 70-kDa protein. Computer analyses of the amino acid sequence predicted patterns of transmembrane domains, α-helices, and coiled coils strikingly similar to those found in mammalian retroviral envelope proteins. In addition, a 65-residue, proline-rich domain is characterized by a strong amino acid compositional bias virtually identical to that of the 60-amino acid, proline-rich neutralization domain of the feline leukemia virus surface protein. The assignment of *SIRE-1* to the *copia*/*Ty1* family was confirmed by comparison of the conceptual translation of its reverse transcriptase-like domain with those of other retroelements. This finding suggests the presence of a proretrovirus in a plant genome and is the strongest evidence to date for the existence of a retrovirus-like genome closely related to *copia*/*Ty1* retrotransposons.**

Retroelements are ubiquitous components of bacterial and eukaryotic genomes that employ reverse transcriptase to sponsor their proliferation (1–3). They encompass a diverse collection of genetic elements that include DNA and RNA viruses, fungal mitochondrial plasmids, bacterial retrons, group II introns, and retrotransposons (1, 3). Infectious retroviruses and related, noninfectious retrotransposons are distinguished from other retroelements, including LINE retroposons, by their possession of long terminal repeats (LTR) (1, 3). Although retroviruses and integrated, endogenous retroviruses are primarily associated with mammalian genomes (2, 4), mammalian LTR retrotransposons have yet to be reported. LTR retrotransposons have been identified in the genomes of other vertebrates (5–8) and are routinely found in the genomes of lower animals, plants, and fungi (1, 3, 9–13). Based on sequence comparisons, some endogenous retroviruses have been shown to be closely related to known infectious retroviruses, whereas others are clearly retrovirus-like but do not correspond to any known infectious viruses (4). In addition to LTR, these retroelements are characterized by genes coding for structural core proteins (*gag*) and four enzymes: protease (*prot*), reverse transcriptase (*rt*), ribonuclease H (*rh*), and integrase (*int*) (1–3). Retroviral genomes encode an envelope protein that mediates both virion export from and entry into susceptible host cells (2, 14).

Most of the characterized LTR retrotransposons belong to either the *copia*/*Ty1* or *gypsy*/*Ty3* group (1). The two classes can be phylogenetically distinguished by amino acid comparisons of the catalytic proteins (1, 15, 16) and by the order of the loci in *pol* (Fig. 1). In all *copia*/*Ty1* elements, *int* precedes *rt* and *rh*, whereas in *gypsy*/*Ty3* group members, *int* resides at the 3′ end of *pol*. All vertebrate retroviruses and endogenous retroviruses conform to the latter configuration (1, 3), and phylogenetic analyses of the conserved regions within the reverse transcriptase suggest that retroviruses and *gypsy*/*Ty3* retroelements are monophyletic (1).

The coding sequences of many characterized plant retrotransposons and endogenous retroviruses are cluttered with disabling stop codons, frameshifts, and deletions, and appear to be nonfunctional (4, 12). Vestiges of ancient *copia*/*Ty1*-like sequences have been identified adjacent to several plant genes (17), and the maize genome apparently contains large clusters of nested retrotransposons (18).

Besides vertebrate retroviruses, five invertebrate *gypsy*/*Ty3* class retroelements (19–23) and a sixth retroelement from a parasitic nematode (24) encode an envelope-like protein. Of these, *gypsy* from *Drosophila melanogaster* (25, 26), Tom from *Drosophila ananassae* (22), and TED from the nocturnal moth, *Trichoplusia ni* (27) produced the encoded protein. Horizontal, infectious-like transfer has been reported for *gypsy* particles (25, 26).

SIRE-1 is a relatively homogeneous family of *copia*/*Ty1*-related retroelements in the soybean genome (28, 29). Each of the several hundred copies is about 11 kb in length (ref. 28; H.M.L., unpublished data), making *SIRE-1* one of the largest retroelements. The family was initially identified after a short segment was fortuitously amplified by the PCR and sequenced (28). We subsequently recovered and sequenced a 2.4-kb cDNA that encompassed the 3′ end of the 5′ LTR, a primer binding site complementary to *Glycine max* $tRNA_{i\text{-}met}$, and an uninterrupted ORF whose conceptual translation produced a retroelement-like, gag-prot polyprotein (29). We now report the characterization of part of a genomic clone that confirms *SIRE-1*'s assignment to the *copia*/*Ty1* family and contains an unprecedented ORF between the 3′ end of the *pol* ORF and the 3′ LTR. The full *SIRE-1* sequence will be published elsewhere.

MATERIALS AND METHODS

Cloning and Sequencing. A λFIXII soybean genomic library (Stratagene) was probed with radiolabeled copies of the *SIRE-1 gag* region as described (28, 29). Positive plaques were purified (30), and DNA from clones carrying the largest inserts were digested with several restriction enzymes. The DNAs were separated by agarose gel electrophoresis and a Southern blot was then probed with an end-labeled, LTR-specific oligonucleotide (30). To isolate possible full-size *SIRE-1* inserts,

This paper was submitted directly (Track II) to the *Proceedings* office.
Abbreviations: LTR, long terminal repeat; rt, reverse transcriptase; int, integrase; rh, ribonuclease H; env, envelope; SU, surface protein; TM, transmembrane protein.
Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. U96295 and AF053008).
*To whom reprint requests should be addressed. e-mail: hlaten@orion.it.luc.edu.

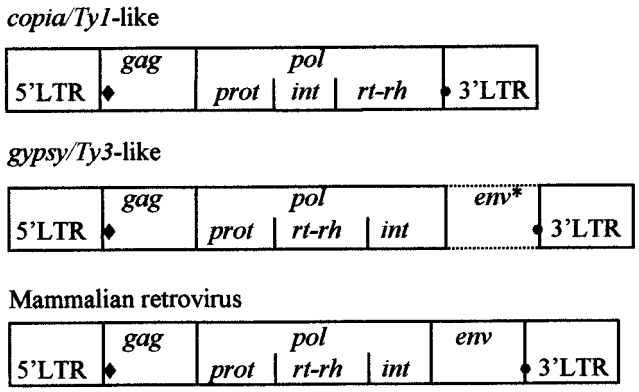

FIG. 1. Organization of LTR retroelement genes. ◆, tRNA primer binding site; ●, polypurine tract. *, Some *gypsy/Ty3* group members possess an *env*-like ORF.

clones in which the probe hybridized to two fragments were selected. DNA was isolated (31) from one phage candidate and digested with *Xba*I and *Hin*dIII. The fragments were then subcloned (30) into pSPORT1 (Life Technologies, Gaithersburg, MD) for automated DNA sequencing. Some longer subclones were unstable, most probably because of rearrangements sponsored by the long direct repeats. Two contiguous, stable subclones, one of which hybridized to the LTR probe, were sequenced on Applied Biosystems Prism 377 DNA sequencers by using pUC/M13 primers and internal primers synthesized at the Loyola University Macromolecular Analysis Facility.

Sequence Analysis and Database Searches. Sequence alignments and ORF determinations were made by using the Genetics Computer Group package (32). Multiple amino acid sequence alignments with seven conserved rt domains suggested by Xiong and Eickbush (33) were made by using PILEUP (32). Trees were constructed by maximum parsimony or neighbor-joining by using PAUP (34). Predictions of α-helices were made by using four programs (35–38). Predictions of coiled coils were generated by using two programs (39, 40), as were predictions of transmembrane domains (41, 42).

Southern Hybridization Analysis. Cloned and genomic DNAs from *G. max* cv Williams 82 were digested to completion in separate reactions with *Bam*HI and *Eco*RI. The digested DNAs were run on a 0.8% agarose gel, blotted, and hybridized (30) to a *rt*-specific probe. After exposure and film development, the membrane was stripped, reexposed to ensure loss of signal, then reprobed with an ORF2-specific probe. The probes were generated by random primer, ^{32}P-labeling (Amersham) of a PCR-amplified segments derived from the two coding regions.

RESULTS

Isolation and Sequence Analysis of Subclones. A genomic clone containing a possible full-size copy of *SIRE-1* was isolated and subcloned as described above. DNAs from two contiguous subclones, the 3′ member of which hybridized to the LTR probe, were sequenced (GenBank accession nos. AF053008 and U96295).

To identify the LTR, the DNA sequence was aligned with that from the *SIRE-1* cDNA clone (29) containing the last 178 bp of a 5′ LTR. The analysis fixed the location of the 3′ end of the LTR on the genomic clone, beyond which the two sequences were unrelated, indicating that the genomic sequence was a 3′ LTR. The genomic and cDNA sequences differed at only four positions (98% identity) over the 178 bp (see Fig. 2).

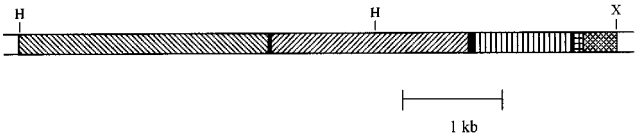

FIG. 2. Organization of *SIRE-1* subclones. ▧, ORF1 (*pol*); ▨, ORF2; ▥, 3′ LTR; ▦, cDNA overlap; ▩, flanking ORF. H, *Hin*dIII; X, *Xba*I.

An uninterrupted, 178-codon ORF adjacent to the 3′ end of the LTR extended to the 3′ end of the 3′ subclone (Fig. 2). This ORF was in the same orientation as the element. Database searches against this ORF by using either the DNA sequence (BLASTN) or the conceptual peptide sequence (BLASTP) did not retrieve any similar sequences. This ORF is presumably the downstream portion of an uncharacterized *G. max* gene split by the *SIRE-1* insertion.

Translation of the remaining DNA produced two ORFs (Fig. 2). ORF1 extended 2,505 bp from the 5′ end of the 5′ subclone. BLASTP searches with the conceptual translation of this sequence retrieved the pol regions of several *copia/Ty1*-like retrotransposons. The alignments demonstrated that, in addition to the 3′ LTR, the subclones encompassed the int, rt, and rh domains of *SIRE-1* (data not shown).

In all *copia/Ty1*-like retrotransposons, *rh* is at the 3′ end of *pol* and is closely followed by a polypurine tract and the 3′ LTR (see Fig. 1). However, the *rh* in *SIRE-1* is followed by a long ORF in the region corresponding to retroviral envelope (*env*) genes (Fig. 2). ORF2 is immediately preceded by a TAA triplet and commences with a threonine codon 27 nt beyond the *pol* stop codon. ORF2 is therefore in the same reading frame as ORF1. Translation of ORF2 would require readthrough of the two stop codons or, alternatively, could be translated as the 3′ member of a spliced transcript (see below).

To confirm the assignment of *SIRE-1* to the *copia/Ty1* family, the conceptual translation of *pol* was aligned to seven conserved retroelement rt domains defined by Xiong and Eickbush (33). The alignments of the second domain are shown in Fig. 3. The aligned sequences were used to build phylogenetic trees by using maximum parsimony (Fig. 4) and

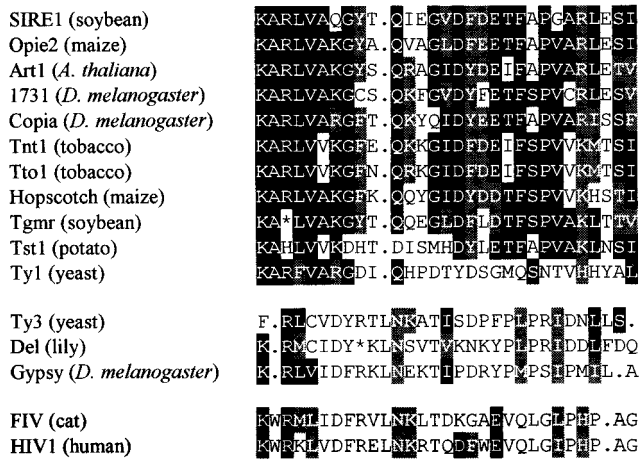

FIG. 3. Multiple sequence alignment of the second conserved domain in rt by using PILEUP (32). *copia/Ty1* consensus positions are highlighted. Amino acids identical to consensus are highlighted in black; amino acids similar to consensus are highlighted in gray. Opie-2 (18); Art1 (C. Herve, unpublished data, GenBank accession no. Y08010); 1731 (43); *Copia* (44); Tnt1 (45); Tto1 (46); Hopscotch (17); Tgmr (47); Tst1 (48); Ty1 (49); Ty3 (50); Del (51); Gypsy (21); FIV (52); HIV1 (53). A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr.

neighbor joining (data not shown) (34). The tree building programs unambiguously placed *SIRE-1* on the *copia/Ty1* branch of the unrooted tree (Fig. 4).

ORF2 is 648 codons in length. The derived theoretical protein has a molecular weight of 70 kDa. Despite its location immediately downstream of *pol*, the translated amino acid sequence (Fig. 5) does not exhibit significant sequence identity to any reported retroviral envelope proteins. This result is not entirely unexpected because known envelope sequences constitute a very heterogeneous collection, and only comparisons between those of closely related retroviruses (e.g., human and simian immunodeficiency viruses, but not human and feline immunodeficiency viruses) reveal recognizable, primary sequence similarities (data not shown). Alternatively, ORF2 could be a transduced cellular sequence. *Bs1* from maize, a low copy-number LTR retrotransposon that lacks its own *rt* (54), contains segments derived from exons of a maize plasma membrane H-ATPase (55, 56).

Identification of Envelope-Like Structural Elements. Retroviral *env* genes encode polypeptides that are cleaved by host proteases into two subunits—surface (SU) and transmembrane (TM) polypeptides—that are subsequently rejoined through disulfide linkages (14, 57). Although the primary sequences of these proteins may be diverse, all retroviral envelope proteins are glycosylated and share three, functionally conserved, hydrophobic transmembrane domains: a signal peptide near the amino terminal of SU (cleaved during processing), a membrane fusion peptide near the amino end of TM, and a distal anchor peptide (14, 57) (Fig. 6).

Retroviral envelope glycoproteins contain between 4 and 30 N-glycosylated asparagines at Asn-Xaa-Ser/Thr motifs (57), with SU generally more heavily glycosylated than TM. The conceptual translation product of ORF2 from *SIRE-1* has only two asparagines in this context. However, retroelement envelope proteins are also known to be O-glycosylated at serine and threonine residues (58, 59). O-glycosylation is correlated with clustering of hydroxy amino acids and elevated frequencies of proline (60). The amino half of the *SIRE-1* theoretical env-like protein conforms to this pattern, and many of the serines and threonines are adjacent to proline. The amino acid composition of one extended, proline-rich region encompassing amino acids 60–127 is similar to the 60-amino acid proline-rich neutralization domain (61) of SU from mammalian leukemia viruses (Table 1). Proline, serine, and threonine are similarly elevated, and there is a nearly complete absence of aromatic amino acids. In *SIRE-1*, the spacing of many of the proline residues—$(Xaa\text{-}Pro\text{-}Yaa)_n$ or $(Xaa\text{-}Pro)_n$—in this region, and from positions 188–197, is characteristic of many structural membrane proteins (62).

The putative env protein sequence was next evaluated for the presence of hydrophobic, membrane-spanning helices (41, 42). Both programs selected the same 13–20 amino acid region centered at residue 30 with high (70–82%) reliability (Fig. 5). The location of the predicted N-terminal, transmembrane helix is consistent with that expected for a signal peptide and is flanked by basic residues, a characteristic feature of most membrane-spanning peptides. Both programs (41, 42) recorded a second transmembrane helix centered at residue 519, but the reliabilities were considerably weaker and of questionable significance. There is, however, a hydrophobic region from residues 510–523 that could correspond to a fusion peptide (Fig. 6, see below).

Only two retroviral env peptides have been structurally characterized by x-ray crystallography (63, 64), but several env SU and TM sequences have been analyzed by structural prediction algorithms (57, 65, 66). Despite the considerable size and sequence diversity among retroviral envelope proteins, these analyses predict multiple α-helical regions similarly distributed throughout the sequence (Fig. 6). The *SIRE-1* envelope-like sequence was evaluated by using several programs whose individual reliabilities ranged from 63% to 70% for predicting short helices in nonhomologous proteins (36–38), to as high as 89% for helices of length greater than eight residues (38). The accuracy of copredicted sites was significantly higher (35). The dispersal of the consensus α-helices predicted in *SIRE-1* resembled that of retroviral proteins (Fig. 5). In addition, the sequence was evaluated for the possible presence of coiled coils (39, 40). Amino acids 580–611 were predicted to form a coiled coil with probabilities approaching 1.0, as were similarly located regions of retroviral TM proteins (Fig. 7). The sequence adheres well to the heptad repeat identified on the carboxyl side of several virus fusion peptides (67–71) (Fig. 8). The predicted coiled coil in the TM domains of HIV and Moloney murine leukemia virus have recently been confirmed by x-ray crystallography (63, 64). Because coiled coils are located near the N terminal in the TM proteins of HIV and the mouse virus, the location of the hydrophobic

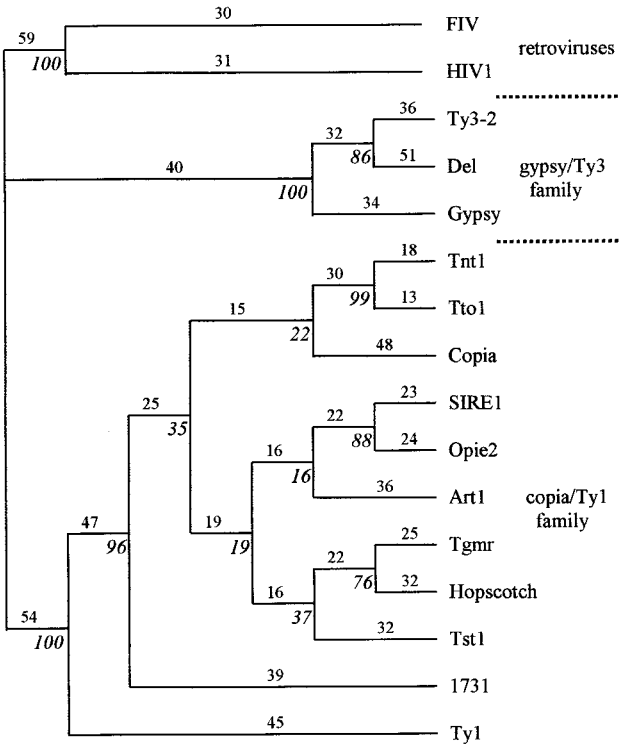

FIG. 4. Maximum parsimony tree based on the alignment of seven conserved domains in *rt*. Numbers above lines are the branch lengths; italicized numbers at nodes are bootstrap values (100 replicates). See Fig. 3 for references.

```
  1  TLIARSLLGQNKFDRCFTRPSTFLIQTHIFVVISFSAFPNSSQRFTKPFQ
 51  RLCFSMATSPKDTSSPGSPSVPSSPSSTKAPSNQEQPEFHIQPIQMIPGQ
101  APVPEKLVPKRQQGVKISENPSIATSPREVDTEMDKKIRSIVSSILKNAS
151  VPDADKDVPTSSTPNAEVLSSSSKEESTEEEQATEETPAPRAPEPAPGD
201  LIDLEEVESDEEPIANKLAPGIAERLQSRKGKTPITRSGRIKTMAQKKST
251  PITPTTSRWSKVAIPSKKRKEFSSSDSDDDVELDVPDIKRAKKSGKKVPG
301  NVPDAPLDNISFHSIGNVERWKFVYQRRLALERELGRDALDCKEIMDLIK
351  AAGLLKTVTKLGDCYESLVREFIVNIPSDITNRKSDEYQKVFVRGKCVRF
401  SPAVINKYLGRPTEGVVDIAVSEHQIAKEITAKQVQHWPKKGKLSAGKLS
451  VKYAILHRIGAANWVPTNHTSTVATGLGKFLYAVGTKSKFNFGKYIFDQT
501  VKHSESFAVKLPIAFPTVLCGIMLSQHPNILNNIDSVMKRESALSLHYKL
551  FEGTHVPDIVSTSGKAAASGAVSKDALIAELKDTCKVLEATIKATTEKKM
601  ELERLIKRLSDSGIDDGEAAEEEEEAAEEEKDAAEDTESDDDDSDATP
```

FIG. 5. Conceptual translation of ORF2. Single underline, predicted transmembrane helix (41, 42); double underline, predicted coiled coil (39, 40); dotted underline, proline-rich region; bold, consensus of predicted α-helices (35–38); wavy underline, possible fusion peptide. ORFs were generated as described (32). See Fig. 3 for amino acid designations.

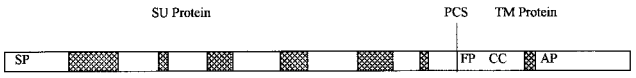

FIG. 6. Predicted and empirically deduced secondary structure features of retroviral envelope proteins (adapted from refs. 65 and 66). ▨, Predicted α-helices; SP, signal peptide; FP, fusion peptide; CC, coiled coil; AP, anchor peptide; PCS, peptide cleavage site.

peptide beginning at residue 511 of the *SIRE-1* ORF2 (Fig. 4) is appropriate for that of a fusion peptide.

Comparison of Cloned and Chromosomal *SIRE-1* Copies. To confirm that the *env*-like gene was not a library or cloning artifact and that its relative location was representative of most, if not all, chromosomal copies of *SIRE-1*, genomic DNA was digested with restriction enzymes, and a Southern blot was sequentially probed with sequences from *rt* and ORF2. Fig. 9*a* shows the positions of the restriction sites relative to the *SIRE-1* coding regions and the probes. As shown in Fig. 9 *b* and *c*, the *rt* and *env*-like probes annealed to the same 4.6-kb *Bam*HI fragment in both the cloned and chromosomal DNAs, confirming that *rt* and the putative *env* are identically juxtaposed in the soybean genome and the clone. The *Eco*RI pattern is more complex. The *rt* probe (Fig. 9*b*), which spans the second *Eco*RI site (see Fig. 9*a*), hybridized with the expected fragments at 1.7 and 0.83 kb in both DNAs. However, there are additional bands in the genomic lane, suggesting that the *Eco*RI sites are polymorphic. The weak upper bands in the clone lanes of Fig. 9*b* are caused by the presence of low levels of vector DNA in the probe.

The *env*-like probe (Fig. 9*c*) hybridized with two unresolved bands in both cloned and genomic DNAs: the same 0.83-kb *Eco*RI fragment that hybridized with the *rt* probe and a 0.85-kb fragment representing the distal *Eco*RI fragment that overlaps the ORF2 probe. The *env*-like probe also hybridized weakly with some of the same putative polymorphic *Eco*RI bands observed with the *rt* probe. Unlike the upper bands in the clone lanes visualized with the *rt* probe, the *env*-like labeled bands in the same lanes are much more intense and are caused by a second copy of this sequence in the original λ clone. The λ*SIRE-1* clone actually contains one complete copy of *SIRE-1* and part of a second copy (H.M.L., unpublished data). The presence of the truncated copy in the upper hybridizing bands was confirmed by a Southern blot of a *Ban*II digest (data not shown). The two copies are not contiguous. We do not know whether the duplication is a cloning artifact or reflects a clustering of elements in the genome, as observed in maize (18).

DISCUSSION

Our data support the inference that *SIRE-1* is an endogenous retrovirus closely related to *copia/Ty1* retrotransposons. All previously characterized retroviruses and endogenous retroviruses are more closely related to *gypsy/Ty3*-like retroelements (Fig. 1). The possibility that in addition to the *gypsy/Ty3* group, some *copia/Ty1* members may actually be endogenous retroviruses suggests that retroviruses have evolved at least twice. The tree in Fig. 4 shows that *SIRE-1* is unequivocally anchored in the *copia/Ty1* family and is most closely related to *opie-2* from maize. The bootstrap values for many of the nodes of the very similar tree generated by neighbor joining (data not shown) were 20–30% higher than the corresponding nodes of the more conservative maximum parsimony tree (Fig. 4). Internal nodes with less than 50% bootstrap support that were consistent with the consensus tree are shown, but their inclusion does not weaken the assignment of *SIRE-1* to the *copia/Ty1* family. *SIRE-1* is not closely related to two other *copia/Ty1*-like soybean retroelements that have been fully (47) or partially (11) characterized. The former, Tgmr (47), is included on the tree in Fig. 4.

The predicted structural features of the ORF2 conceptual translation product are similar to those found in retroviral envelope proteins. The correspondence of conserved features is not perfect, however. The *SIRE-1* envelope-like sequence has fewer glycosylation sites and appears to be missing a transmembrane anchor peptide. In addition, the *SIRE-1* sequence has far fewer cysteine residues, some of which sponsor disulfide bridges within and between SU and TM (57). Retroviral envelope proteins are generated from spliced transcripts (2, 57). In the case of some avian retroviruses, splicing leads to an in-frame fusion of the *gag* start codon with the 5′ end of *env* (57), obviating the need for an initiation codon in *env*. An analogous splice in a *SIRE-1* transcript would serve the same purpose, although no splice donor or acceptor consensus sequences were found in the expected regions. Cleavage of mammalian retroviral envelope precursors into SU and TM generally occurs at a conserved site near the amino terminal of the fusion peptide at the consensus (Arg/Lys)-Xaa-(Arg/Lys)-Arg (57). This sequence does not appear in the putative *SIRE-1* envelope protein, and the only appropriately located tetrapeptide with at least two basic amino acids is at position 487. Complete adherence to the full catalog of generally conserved, retroviral envelope features, however, should not be expected because it is unlikely that the *SIRE-1* and retroviral *env* genes are related by descent. Phylogenetic analyses suggest that the *copia/Ty1* and *gypsy/Ty3* groups diverged from each other prior to the emergence of enveloped retroviruses from the *gypsy/Ty3* line of descent (1, 15, 33).

In addition to demonstrating the congruence of the cloned and chromosomal copies of *SIRE-1*, the comparable intensities of the hybridization signals between the clone and chromosomal lanes in Fig. 9 *b* and *c*, attest to the high copy number of "env"-containing *SIRE-1* members. Although the possibility cannot be ruled out, we do not believe this *env*-like ORF is a transduced host gene. The presence in retrotransposons of apparently transduced host genes has been found only rarely (12, 55, 56). The maize *Bs1* element appears to have sacrificed its *rt* gene to gain a cellular sequence and is apparently not capable of autonomous retrotransposition (54). The presence of the *env*-like ORF in most if not all of the several hundred copies of *SIRE-1* suggests that this gene is an integral part of a retroelement genome that was, or is, functional, at least as a retrotransposon. Preliminary sequence analysis of regions upstream of *int* (H.M.L. and E. Gaucher, unpublished data) coupled with the previously characterized cDNA clone (29) indicate that ORF1 also encompasses *gag* and *prot* regions of appropriate length.

Neither retroviral genomes nor virions have been reported in plants, although both classes of retrotransposons are widespread. Plant caulimoviruses encode reverse transcriptase, but

Table 1. Comparison of 60-residue proline-rich regions from *SIRE-1* and mammalian retroviruses

	Amino acid composition, %										
Element	P	S+T	F+W+Y	I+L+V	H	N	Q	A+G	D+E	K+R	C+M
SIRE-1	20	23	1	15	2	3	11	11	6	8	1
FeLV	18	22	0	20	3	5	8	13	3	5	2
MLV	28	18	0	20	0	2	6	12	4	10	0

FeLV, feline leukemia virus; MLV, murine leukemia virus. See Fig. 3 for amino acid abbreviations.

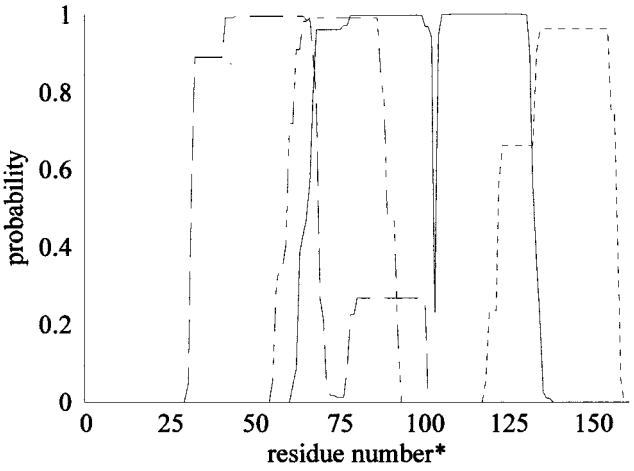

FIG. 7. Output of COILS program search for coiled coils (39) in *SIRE-1* and selected retroviral transmembrane proteins. Similar results were obtained with ref. 40. Results shown are for a window of 21 amino acids. ——, *SIRE-1*; · · ·, HIV; - - -, murine leukemia virus; ·····, influenza A virus. *, Numbering from the peptide cleavage site, except for *SIRE-1*, where residue 1 corresponds to the amino end of the putative hydrophobic fusion peptide.

have DNA genomes and do not integrate into host chromosomes (72). Very few plant virus genomes encode an *env* gene. Those that do—rhabdoviruses and bunyaviruses (72)—also infect animal hosts, where envelope proteins sponsor viral-host cell membrane fusion. Intact plant cell walls may hinder this mode of virus transfer, and whether viral envelope proteins serve the same function in plant hosts as they do in animals is not known. This finding suggests that *SIRE-1* may have originally been an infectious invertebrate retrovirus that was transferred to soybean by the invertebrate vector. In plants, intercellular virus spread is mediated by movement proteins (73), but there is no evidence for the existence of this property in any of the theoretical *SIRE-1* gene products.

Most higher plant retrotransposons with copy numbers comparable to *SIRE-1* are very heterogeneous and are composed of multiple subfamilies (10, 74–77) analogous to retroviral quasispecies (78). The absence of additional hybridizing bands in the chromosomal lanes in Fig. 9 *a* and *b* is therefore unusual. Genomic digests of soybean DNA generated by a dozen different restriction enzymes have now been probed with cloned copies of the *gag*-like, *rt*-like, and *env*-like regions of *SIRE-1*. Subfamilies were not detected in any of these digests, although a few low copy-number derivatives may be present (Fig. 9; also ref. 28 and H.M.L., unpublished data). This general, restriction-site homogeneity, the presence of long, uninterrupted ORFs within and adjacent to *SIRE-1*, and the near identities of the comparable 178 bp of the two LTRs suggest that the introduction and amplification of *SIRE-1* in *G. max* and its wild progenitor, *Glycine soja*, is a relatively recent event. Functional copies of *SIRE-1* may persist. Transcripts containing *gag*-like, *rt*-like, and *env*-like sequences have been detected by Northern blot hybridization and RT-PCR, and it appears that the 5′ end of some, if not all of these are located within the LTR (E. Lin and H.M.L., unpublished data).

		a d a d a d a d a d a d a
HIV (53)	545	LSGIVQQQNNLLRALEASYAMVQNIAKGIRILEARVARVEALV
MLV (70)	441	TQQFQQLQAAVQDDLREVEKSISNLEKSLTSLSEVVLQNRRGL
InflA (71)	419	KYVEDKILWSYNAELLVALENQHTIDLTDSEMNKLFEKTRRQL
SIRE-1	577	IAELKDTCKVLEATIKATTEKKMELERLIKRLSDSGIDDGEAA

FIG. 8. Heptad repeats in the coiled coil region of retroviral TM proteins. MLV, murine leukemia virus; InflA: influenza virus type A. See Fig. 3 for amino acid designations.

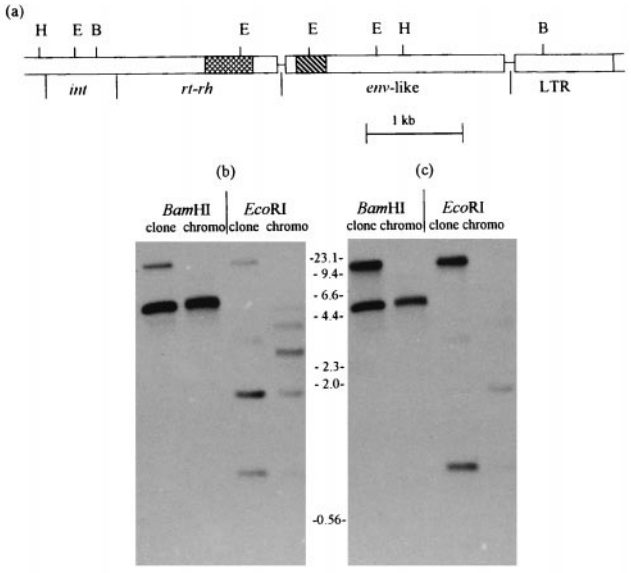

FIG. 9. Structural congruence of *pol* and *env*-like regions of *SIRE-1* from a λ clone and chromosomal copies. (*a*) Restriction map showing *Hin*dIII (H), *Eco*RI (E), and *Bam*HI (B) sites relative to *rt*- (▨) and *env*- (▧) specific probes. *Bam*HI and *Eco*RI digests of clone and chromosomal (chromo) DNAs sequentially hybridized with *rt* (*b*) and ORF2 (*c*) probes

The occurrence of genetically related retrotransposons among phylogenetically unrelated hosts has led to the assumption that these noninfectious elements can be transferred horizontally by some unknown mechanism (1). The observation that members of both major LTR retrotransposon families have *env*-like genes provides a foundation for the counterintuitive proposal that the apparent horizontal transfer of LTR retrotransposons may be the result of transmission of closely related retroviral derivatives that subsequently lost their *env* gene. Although many dozens of presumed *copia*/*Ty1*-related retrotransposons have been detected by PCR amplification of conserved *rt* domains (5, 10, 11, 74–77), the number of fully sequenced elements is relatively small (1, 3, 12). It is conceivable that additional *env* regions may be encountered as more of these elements are fully sequenced.

We thank Z. Burki and C. Brown for their help with automated DNA sequencing, W. Ballard and J. Norman for their assistance with tree building, and S. Wessler for advice and comments on the manuscript. The assistance and contributions of Y.-A. Bi, M. Hughes, M. Grassi, A. Sverdlik, and S. Wakim are gratefully acknowledged. We are also grateful to the scientists and organizations that made their analytical resources available on the World Wide Web.

1. Eickbush, T. H. (1994) in *The Evolutionary Biology of Viruses*, ed. Morse, S. S. (Raven, New York), pp. 121–157.
2. Varmus, H. & Brown, P. (1989) in *Mobile DNA*, eds. Berg, D. E. & Howe, M. M. (Am. Soc. Microbiol., Washington, DC), pp. 53–108.
3. Flavell, A. J. (1995) *Comp. Biochem. Physiol. B* **110,** 3–15.
4. Urnovitz, H. B. & Murphy, W. H. (1996) *Clin. Microbiol. Rev.* **9,** 72–99.
5. Flavell, A. J., Jackson, V., Iqbal, M. P., Riach, I. & Waddell, S. (1995) *Mol. Gen. Genet.* **246,** 65–71.
6. Flavell, A. J. & Smith, D. B. (1992) *Mol. Gen. Genet.* **233,** 322–326.
7. Tristem, M., Kabat, P., Herniou, E., Karpas, A. & Hill, F. (1995) *Mol. Gen. Genet.* **249,** 229–236.
8. Greene, J. M., Otain, H., Good, P. J. & Dawid, I. B. (1993) *Nucleic Acids Res.* **21,** 2375–2381.
9. Britten, R. J., McCormack, T. J., Mears, T. L. & Davidson, E. H. (1995) *J. Mol. Evol.* **40,** 13–24.
10. Flavell, A. J., Smith, D. B. & Kumar, A. (1992) *Mol. Gen. Genet.* **231,** 233–242.

11. Voytas, D. F., Cummings, M. P., Konieczny, A., Ausubel, F. M. & Rodermel, S. R. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 7124–7128.
12. Bennetzen, J. L. (1996) *Trends Microbiol.* **4,** 347–353.
13. Britten, R. J. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 599–601.
14. Freed, E. O. & Martin, M. A. (1995) *J. Biol. Chem.* **270,** 23883–23886.
15. Doolittle, R. F., Feng, D.-F., Johnson, M. S. & McClure, M. A. (1989) *Q. Rev. Biol.* **64,** 1–30.
16. McClure, M. A. (1991) *Mol. Biol. Evol.* **8,** 835–856.
17. White, SE, Habera, L. F. & Wessler, S. R. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 11792–11796.
18. SanMiguel, P., Tikhonov, A., Jin, Y.-K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P. S., Edwards, K. J., Lee, M., Avramova, Z., *et al.* (1996) *Science* **274,** 765–768.
19. Saigo, K., Kugiyama, W., Matsuo, Y., Inouye, S., Yoshioka, K. & Yuki, S. (1984) *Nature (London)* **312,** 659–661.
20. Inouye, S., Yuki, S. & Saigo, K. (1986) *Eur. J. Biochem.* **154,** 417–425.
21. Marlor, R. L., Parkhurst, S. M. & Corces, V. G. (1986) *Mol. Cell. Biol.* **6,** 1129–1134.
22. Tanda, S., Mullor, J. L. & Corces, V. G. (1994) *Mol. Cell. Biol.* **14,** 5392–5401.
23. Friesen, P. D. & Nissen, M. S. (1990) *Mol. Cell. Biol.* **10,** 3067–3077.
24. Felder, H., Herzceg, A., deChastonay, Y., Aeby, P., Tobler, H. & Muller, F. (1994) *Gene* **149,** 219–225.
25. Song, S. U., Gerasimova, T., Kurkulos, M., Boeke, J. D. & Corces, V. G. (1994) *Genes Dev.* **8,** 2046–2057.
26. Kim, A., Terzian, C., Santamaria, P., Pelisson, A., Prud'homme, N. & Bucheton, A. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 1285–1289.
27. Ozers, M. S. & Friesen, P. D. (1996) *Virology* **226,** 252–259.
28. Laten, H. M. & Morris, R. O. (1993) *Gene* **133,** 153–159.
29. Bi, Y.-A. & Laten, H. M. (1996) *Plant Mol. Biol.* **30,** 1315–1319.
30. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY).
31. Burmeister, M. & Lehrach, H. (1996) *Trends Genet.* **12,** 389.
32. Devereux, J., Haeberli, P. & Smithies, O. (1984) *Nucleic Acids Res.* **12,** 387–395.
33. Xiong, Y. & Eickbush, T. H. (1990) *EMBO J.* **9,** 3353–3362.
34. Swofford, D. L. (1997) PAUP (Sinauer Associates, Sunderland, MA).
35. Geourjon, C. & Deleage, G. (1995) *Comput. Appl. Biosci.* **11,** 681–684.
36. Gibrat, J. F., Garnier, J. & Robson, B. (1987) *J. Mol. Biol.* **198,** 425–443.
37. Levin, J. M., Robson, B. & Garnier, J. (1986) *FEBS Lett.* **205,** 303–308.
38. Solamov, A. A. & Solovyev, V. V. (1995) *J. Mol. Biol.* **247,** 11–15.
39. Lupas, A., Dyke, M. & Van Stock, J. (1991) *Science* **252,** 1162–1164.
40. Wolf, E., Kim, P. S. & Berger, B. (1997) *Protein Sci.* **6,** 1179–1189.
41. Hofmann, K. & Stoffel, W. (1993) *Biol. Chem. Hoppe-Seyler* **374,** 166.
42. Rost, B., Casadia, R., Fariselli, P. & Sander, C. (1995) *Protein Sci.* **4,** 521–533.
43. Fourcade-Peronnet, F., d'Auriol, L., Becker, J., Galibert, F. & Best-Belpomme, M. (1988) *Nucleic Acids Res.* **16,** 6113–6125.
44. Mount, S. M. & Rubin, G. M. (1985) *Mol. Cell. Biol.* **5,** 1630–1638.
45. Grandbastien, M.-A., Spielmann, A. & Cabouche, M. (1989) *Nature (London)* **347,** 376–380.
46. Hirochika, H., Otsuki, H., Yoshikawa, M., Otsuki, Y. & Sugimoto, K. (1996) *Plant Cell* **8,** 725–734.
47. Bhattacharyya, M. K., Gonzales, R. A., Kraft, M. & Buzzell, R. I. (1997) *Plant Mol. Biol.* **34,** 255–264.
48. Camirand, A., St-Pierre, B, Marineau, C. & Brisson, N, (1990) *Mol. Gen. Genet.* **224,** 33–39.
49. Clare, J. & Farabaugh, P. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 2829–2833.
50. Hansen, L. J., Chalker, D. L. & Sandmeyer, S. B. (1988) *Mol. Cell. Biol.* **8,** 5245–5256.
51. Smyth, D. R., Kalitsis, P., Joseph, J. L. & Sentry, J. W. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 5015–5019.
52. Olmsted, R. A., Hirsch, V. M., Purcell, R. H. & Johnson, P. R. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 8088–8092.
53. Ratner, L., Haseltine, W., Patearca, R., Livak, K. J., Starcich, B. R., *et al.* (1985) *Nature (London)* **313,** 277–284.
54. Jin, Y.-K. & Bennetzen, J. L. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 6235–6239.
55. Bureau, T. E., White, SE & Wessler, S. R. (1994) *Cell* **77,** 479–480.
56. Palmgren, M. G. (1994) *Plant Mol. Biol.* **25,** 137–140.
57. Hunter, E. & Swanstrom, R. (1990) *Curr. Top. Microbiol. Immunol.* **157,** 187–253.
58. Pinter, A. & Honnen, W. J. (1988) *J. Virol.* **62,** 1016–1021.
59. Bernstein, H. B., Tucker, S. P., Kar, S. R., McPherson, S. A., McPherson, D. T., Dubay, J. W., Lebowitz, J., Compans, R. W. & Hunter, E. (1995) *J. Virol.* **69,** 2745–2750.
60. Wilson, I. B. H., Gavel, Y. & von Heijne, G. (1991) *Biochem. J.* **275,** 529–534.
61. Fontenot, J. D., Tjandra, N., Ho, C., Andrews, P. C. & Montelaro, R. C. (1994) *J. Biomol. Struct. Dyn.* **11,** 821–836.
62. Williamson, M. P. (1994) *Biochem. J.* **297,** 249–260.
63. Chan, D. C., Fass, D., Berger, J. M. & Kim, P. S. (1997) *Cell* **89,** 263–273.
64. Fass, D., Harrison, S. C. & Kim, P. S. (1996) *Nat. Struct. Biol.* **3,** 465–469.
65. Gallaher, W. R., Ball, J. M., Garry, R. F., Martin-Amedee, A. M. & Montelaro, R. C. (1995) *AIDS Res. Hum. Retroviruses* **11,** 191–202.
66. Gallaher, W. R., Ball, J. M., Garry, R. F., Griffin, M. C. & Montelaro, R. C. (1989) *AIDS Res. Hum. Retroviruses* **5,** 431–440.
67. Chambers, P., Pringle, C. R. & Easton, A. J. (1990) *J. Gen. Virol.* **71,** 3075–3080.
68. Rabenstein, M. & Shin, Y.-K. (1995) *Biochemistry* **34,** 13390–13397.
69. Hughson, F. M. (1995) *Curr. Biol.* **5,** 265–274.
70. Shinnick, T. M., Lerner, R. A. & Sutcliffe, J. G. (1981) *Nature (London)* **293,** 543–548.
71. Gething, M. J., Bye, J., Skehel, J. & Waterfield, M. (1980) *Nature (London)* **287,** 301–306.
72. Matthews, R. E. F. (1991) *Plant Virology* (Academic, New York).
73. Mushegian, A. R. & Koonin, E. V. (1993) *Arch. Virol.* **133,** 239–257.
74. Flavell, A. J., Dunbar, E., Anderson, R., Pearce, S. R., Hartley, R. & Kumar, A. (1992) *Nucleic Acids Res.* **14,** 3639–3644.
75. VanderWiel, P. L., Voytas, D. F. & Wendel, J. F. (1993) *J. Mol. Evol.* **36,** 429–447.
76. Pearce, S. R., Harrison, G., Li, D., Heslop-Harrison, J. S., Kumar, A. & Flavell, A. J. (1996) *Mol. Gen. Genet.* **250,** 305–315.
77. Wang, S., Zhang, Q., Maughan, P. J. & Saghai Maroof, M. A. (1997) *Plant Mol. Biol.* **33,** 1051–1058.
78. Holland, J. J., de la Torre, J. C. & Steinhauer, D. A. (1992) *Curr. Top. Microbiol. Immunol.* **176,** 1–20.