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

Edited 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 retroelements, 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 (prr), reverse transcriptase (rt), ribonuclease H (rhn), 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 r 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 tRNAi-met, and an uninterrupted ORF whose conceptual translation produced a retroviral envelope-like protein, 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 AFIXII 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.

© 1998 by The National Academy of Sciences 0027-8424/98/956897-652.00/0 PNAS is available online at http://www.pnas.org.

© 1998 by The National Academy of Sciences 0027-8424/98/956897-652.00/0 PNAS is available online at http://www.pnas.org.
clones in which the probe hybridized to two fragments were selected. DNA was isolated (31) from one phage candidate and digested with XhoI and HindIII. 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 Eckbush (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 BamHI and EcoRI. 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 (30), the membrane was stripped, reexposed to ensure loss of signal, then reprobed with an ORF2-specific probe. The probes were generated by random primer, 32P-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).

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 (Fig. 2) 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. BLAST searches with the conceptual translation of this sequence retrieved the pol regions of several copia/Tyl-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/Tyl-like retrotransposons, rh is at the 3’ end of pol and is closely followed by a polyurine 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/Tyl family, the conceptual translation of pol was aligned to seven conserved retroelement rt domains defined by Xiong and Eckbush (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. 1. Organization of LTR retroelement genes.](https://example.com/fig1.png)

![Fig. 2. Organization of SIRE-1 subclones.](https://example.com/fig2.png)

![Fig. 3. Multiple sequence alignment of the second conserved domain in rt by using PILEUP (32).](https://example.com/fig3.png)
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. Bs f 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-Pro-Yaa), or (Xaa-Pro)—, 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

1. T1IARSLSLLQGKNDKCFCRSTPFRSTFLQCHIFWV182AFNFSNQRTKFDQ
2. RLCSFMTSPKTSSPGPSPSFBFSTKMPNQEEFHP1QIQ1KQ
3. APFFVKFLKQGGLIIASELPATFV4DETDKIKTSVVLSNRAS
4. VEPASKDGYTPSSTPAVLV3SESSKSTREVQQSARPFFPSERPGD
5. LIDTEEVSSDEPIKAPLAIERIQ8NSRKTIPITRSGKNIKTAKQKST
6. 2PITPTTSMRSKVALIP5REKKEFSSDSS DOVEDLOVDIKRAKKGKVSFG
7. NPVDPLNR15SHI GNERWKFVFQQLARLERLORDALC OME1L
8. 3AQL1ATLYVTEGRLCESLHIWPI20NIPIDTINERKGQVFGYCGCRF
9. SAVAKLYGLPRTDQVISATMSQ14ATKQ1Q9WHPKEKGLSAGKLS
10. VTXAILHRIGAAGDVPVTNMT7STVLGTGKLFLAYVCTKSNFKFGY 1FLG
11. VNRSFENKFPFAPFLQGSEARCHT185QNLDSMRSSALRSEYFL
12. FEGTFYPEIVF20T3SGRAARAGVSVDGAELCQ4PVV3SSATAT11001
13. EELERLLKSDQIDGDDPSAEKAEEASAEDAEETSDEDDSSDATT

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.

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.
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 sequenced and hybridized with sequence from the clone. The ORF2 probe. Unlike the upper bands in the clone lane visualized with the EcoRI probe, the ORF2 probe also hybridized weakly with some of the same putative polymorphic bands in the genomic lane, suggesting that the EcoRI sites are polymorphic. The weak upper bands in the clone lanes of Fig. 9b and c, the rt and env-like probes annealed to the same 4.6-kb BamHI fragment in both the cloned and chromosomal DNAs, confirming that the sites are identical and exposed in the soybean genome and the clone. The EcoRI pattern is more complex. The rt probe (Fig. 9b), which spans the second EcoRI site (see Fig. 9a), 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 EcoRI sites are polymorphic. The weak upper bands in the clone lanes of Fig. 9b are caused by the presence of low levels of vector DNA in the probe.

The env-like probe (Fig. 9c), hybridized with two unresolved bands in both cloned and genomic DNAs: the same 0.83-kb EcoRI fragment that hybridized with the rt probe and a 0.85-kb fragment representing the distal EcoRI site that overlaps the ORF2 probe. The env-like probe also hybridized weakly with some of the same putative polymorphic EcoRI 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 λSIRE-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 BanII 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. 9b 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 pol 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

<table>
<thead>
<tr>
<th>Element</th>
<th>P</th>
<th>S+T</th>
<th>F+W+Y</th>
<th>I+L+V</th>
<th>H</th>
<th>N</th>
<th>O</th>
<th>A+G</th>
<th>D+E</th>
<th>K+R</th>
<th>C+M</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIRE-1</td>
<td>20</td>
<td>23</td>
<td>1</td>
<td>15</td>
<td>2</td>
<td>3</td>
<td>11</td>
<td>11</td>
<td>6</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>FeLV</td>
<td>18</td>
<td>22</td>
<td>0</td>
<td>20</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>13</td>
<td>3</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>MLV</td>
<td>28</td>
<td>18</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>12</td>
<td>4</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

FeLV, feline leukemia virus; MLV, murine leukemia virus. See Fig. 3 for amino acid abbreviations.
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 178bp of the two LTRs are very heterogeneous and are composed of multiple subfamilies (10, 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.

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/Tyl1-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.


Protein Sci. 8, 725–734.


6902 Genetics: Laten et al.  

6084 E. coli K-12 Genetic Mutant Collection  
