Large Retrotransposon Derivatives: Abundant, Conserved but Nonautonomous Retroelements of Barley and Related Genomes

Ruslan Kalendar,* Carlos M. Vicient,* Ofer Peleg,† Kesara Anamthawat-Jonsson,‡ Alexander Bolshoy† and Alan H. Schulman*,§,1

**MTT/BI Plant Genomics Laboratory, Institute of Biotechnology, Viikki Biocenter, University of Helsinki, FIN-00014 Helsinki, Finland, †Genome Diversity Center, Institute of Evolution, University of Haifa, 31905 Haifa, Israel, ‡Faculty of Sciences, University of Iceland, 108 Reykjavík, Iceland and §Plant Breeding Biotechnology, Plant Production Research, MTT Agrifood Research Finland, 31600 Jokioinen, Finland*

Manuscript received August 18, 2003
Accepted for publication November 24, 2003

ABSTRACT

Retroviruses and LTR retrotransposons comprise two long-terminal repeats (LTRs) bounding a central domain that encodes the products needed for reverse transcription, packaging, and integration into the genome. We describe a group of retrotransposons in 13 species and four genera of the grass tribe Triticeae, including barley, with long, ~4.4-kb LTRs formerly called *Sukkula* elements. The ~3.5-kb central domains include reverse transcriptase priming sites and are conserved in sequence but contain no open reading frames encoding typical retrotransposon proteins. However, they specify well-conserved RNA secondary structures. These features describe a novel group of elements, called LARDs or *la*rge *r*etrotransposon *d*erivatives (LARDs). These appear to be members of the *gypsy* class of LTR retrotransposons. Although apparently nonautonomous, LARDs appear to be transcribed and can be recombinationally mapped due to the polymorphism of their insertion sites. They are dispersed throughout the genome in an estimated 1.3×10^3 full-length copies and 1.16×10^4 solo LTRs, indicating frequent recombinational loss of internal domains as demonstrated also for the *BARE*-1 barley retrotransposon.

RETROTRANSPOSONS are ubiquitous in the genomes of plants, animals, and fungi (Flavell *et al.* 1992; Voytas *et al.* 1992; Suoniemi *et al.* 1998a). They replicate by a cycle of transcription, reverse transcription, and integration of the cDNA copies back into the genome (Kumar and Bennetzen 1999). This life cycle, composed of a feedback loop for replicationally active elements, gives retrotransposons the potential to form large fractions of the genome. The LTR retrotransposons have a structure, as do retroviruses, composed of two long terminal repeats (LTRs) at either end and an internal coding domain that specifies the protein products (Frankel and Young 1998; Kumar and Bennetzen 1999) needed for replication, packaging into virus-like particles, and integration into the genome (Figure 1). Between the left and right LTRs, respectively, and the coding domain of these LTR retrotransposons are found the priming sites for minus- and plus-strand cDNA synthesis. Within retrotransposon families, the protein products such as integrase and reverse transcriptase can be highly conserved and show purifying selection for function (McAllister and Werren 1997; Suoniemi *et al.* 1998b). Sequence similarities and coding domain order divide the families of LTR retrotransposons into the *copia*-like and *gypsy*-like classes, which appear to have been distinct since early in eukaryotic evolution.

Despite conservation for function, the processes of transcription and reverse transcription are highly error prone (Gabriel and Mules 1999), contributing to the occurrence of many translationally defective copies in the genome (Suoniemi *et al.* 1998b) and leading to the formation of quasi-species for retrotransposon families (Casacuberta *et al.* 1995). For retroviruses and retrotransposons, retroelement products can functionally complement defective copies (Heidmann *et al.* 1988; Hwang *et al.* 2001). Among the DNA transposons, defective or nonautonomous elements, which can be mobilized by the transposase of intact, autonomous elements, are commonplace (Hartl *et al.* 1992). Recently, retroelements analogous to the nonautonomous DNA transposons were identified as widespread components of plant genomes (Witte *et al.* 2001). These terminal repeat retrotransposons in miniature (TRIM) elements are composed of 100- to 250-bp terminal direct repeats, which appear to be LTRs or LTR derivatives, and the priming motifs normally found internal to the LTRs in full-length retrotransposons and retroviruses. However, aside from the priming sites and a small intervening segment, the TRIM elements almost completely lack internal

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. AF453658–AF453684, AY054376–AY054381, AY069966, AY069967, ALIGN_000280, ALIGN_000282, and ALIGN_000601.

[1]*Corresponding author:* Plant Genomics Laboratory, Institute of Biotechnology, University of Helsinki, P.O. Box 56, Viikinkaari 4, FIN-00014 Helsinki, Finland. E-mail: alan.schulman@helsinki.fi

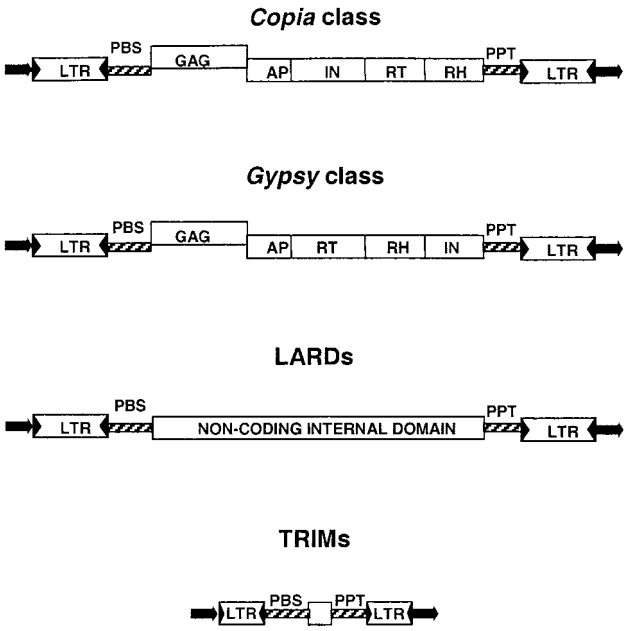

FIGURE 1.—Organization of the major types of LTR retrotransposons. *Copia*-like and *gypsy*-like elements encode capsid proteins (GAG), aspartic proteinase (AP), integrase (IN), reverse transcriptase (RT), and RNase H (RH), translated as a polyprotein. GAG is often in a different reading frame from the other products, shown as a shifted box. The protein-coding region is flanked by the (−)-strand (PBS) and (+)-strand (PPT) priming sites for reverse transcription and by long terminal repeats (LTRs). The LTRs contain terminal inverted repeats (solid triangles). The retroelements are flanked by direct repeats (solid arrows). In LARD elements, the coding region is replaced by a long, conserved noncoding domain whereas in TRIM elements the central domain is almost completely lacking and the LTRs are quite short. Components are not drawn to scale.

domains. The evidence for past mobility of TRIMs in the genome suggests that these may be mobilized through complementation by the protein products of intact retrotransposons.

In the first full-length clone of *BARE*-1 (Z17327), a barley (*Hordeum vulgare*) LTR retrotransposon, we reported a 4930-bp sequence in the 3′ LTR (MANNINEN and SCHULMAN 1993) flanked by 5-bp direct repeats reminiscent of retrotransposon insertions. We later reported two ~5-kb sequences similar to this insertion in a 66-kb stretch of the barley genome (SHIRASU *et al.* 2000), which were also flanked by 5-bp direct repeats, and named this set of sequences *Sukkula*, Finnish for "shuttle." The sequenced *Sukkula* elements resembled solo LTRs, but no full-length elements were identified. However, polymorphisms for *Sukkula* insertion sites in both barley and *Aegilops tauschii* enabled a set of these elements to be recombinationally mapped (MANNINEN *et al.* 2000; BOYKO *et al.* 2002). Here, we have isolated and sequenced *Sukkula*-like elements from barley, other species in the tribe Triticeae, and from species outside the Triticeae. We have also isolated and sequenced intervening domains lying between *Sukkula* elements. The results indicate that *Sukkula* elements are LTRs and that the LTR pairs together with the intervening domains constitute complete, but nonautonomous elements belonging to a novel group of retroelements. Members of this group are similar to TRIMs in their lack of a protein-coding domain. However, they are unique in their possession of a large internal domain that is highly conserved in primary sequence and secondary structure despite its lack of coding capacity. We name this group *la*rge *r*etrotransposon *d*erivative (LARD) elements. We have estimated their copy number, examined their chromosomal distribution, and characterized their structure.

MATERIALS AND METHODS

Plant materials: Seeds were a gift of Dr. Ole Seberg (Botanical Institute, Copenhagen, Denmark) or of Boreal Plant Breeding Limited (Jokioinen, Finland). Plant DNA was isolated from seedling leaves by the CTAB method essentially as previously described (KALENDAR *et al.* 1999).

Bacterial artificial chromosome manipulations: The BAC clones of barley genomic DNA, their handling, and DNA blotting from them were described previously (VICIENT *et al.* 1999). For DNA blots, 500 ng BAC clone DNA was digested with the appropriate enzyme and separated on agarose gels, denatured, and transferred to a Hybond N+ membrane (Amersham Biosciences, Arlington Heights, IL), using standard methods (SAMBROOK *et al.* 1989). The *Sukkula* LARD LTR probe consisted of a PCR fragment amplified from barley (cv. Bomi) genomic DNA with the following primers: Suk5, 5′-CGCTACG GTCGACCGTTCGGGTACC-3′, which matches the original *Sukkula* insertion in *BARE*-1 (Z17327) at nucleotides 7926–7902, and 91673, 5′-TGTGACAGCCCGATGCCGACGTTCC-3′ (Z17327, nucleotides 7558–7582). Probes were random primed (Rediprime II or Megaprime; Amersham Biosciences) and ^{32}P labeled. Membrane prehybridization was carried out in 0.25 M $NaHPO_4$ (pH 7.2), 7% SDS (w/v), and 1 mM EDTA to which 100 mg liter^{-1} of previously denatured herring sperm DNA was added. After 20 min prehybridization at 65°, the DNA probe was added to the prehybridization buffer. Hybridization was performed overnight. The hybridized filters were then washed twice with 2× SSC, 0.1% SDS for 30 min at 65°. Bound radiation was quantified by exposure to an imaging plate as before (VICIENT *et al.* 1999).

Determination of LARD LTR sequences: Primers were designed with the Fast PCR program (http://www.biocenter.helsinki.fi/bi/bare-1_html/oligos.htm). They were made to match the ends of the four LARD LTR, or *Sukkula*, sequences available, two from a 66-kb contiguous sequence at the *Rar1* locus of chromosome 2HL (AF254799, 11,204–6645; 20,145–24,048, 25,780–27,570), the insertion in the 3′ LTR of *BARE*-1a (Z17327, nucleotides 5758–10,687), and one on chromosome 4H near the *Mlo* gene (Y14573, 42,140–37,153). These primers were 91673, described above, and 91674, 5′-CACGC CCAAGATGCGACCCTATCC-3′ (Z17327, nucleotides 10,684–10,661). The PCR was carried out with 20 ng genomic DNA in a 20-μl reaction containing 50 mM Tris-HCl, pH 9.0, 15 mM $(NH_4)_2SO_4$, 1.5 mM $MgCl_2$, 0.1% (v/v) Triton X-100, 200 nM each primer, 200 μM dNTPs, and 1 unit DyNAzymeEXT DNA Polymerase (Finnzymes, Espoo, Finland). The amplification program consisted of denaturation at 94°, 2 min; 15 cycles of 94°, 20 sec, 68°, 4 min; and a final elongation step

at 72°, 10 min. The PCR was performed in a Master Cycler Gradient (Eppendorf, Madison, WI) or PTC-225 DNA Engine Tetrad (MJ Research, Watertown, MA) in 0.2-ml tubes (AB Advanced Biotechnologies, Epsom, United Kingdom). Products were cloned in the pGEM-5Zf(+) T-vector (Promega, Madison, WI). Full-length sequences for the LTR were derived by primer walking in both directions, and sequencing was carried out in house on automated sequencers (http://www.biocenter.helsinki.fi/bi/dna/).

Determination of internal LARD sequences: Long-distance PCR for the cloning of LARD internal domains was carried out as for the LTRs for all species and genera except Triticum and Secale. The 5′ primer, LTROL2, faced outward from the 5′ LTR 280 nt from the terminus and consisted of 5′-GCAG CCTGGGATAGCAAGGATGG-3′. The 3′ primer, LTROL, was located 146 nt from the 5′ terminus of the 3′ LTR and consisted of 5′-CCGGCAGCTACGAACGGATGCAAG-3′. Strong, single bands were obtained, cloned, and sequenced in both directions. For all except Triticum and Secale, the 5′ primer, 9900, consisted of 5′-GATAGGGTCGCATCTTGGGCGTGAC-3′ and was used in combination with primer LTROL.

Determination of LTR copy number: The sources and provenances of barley cv. Bomi and of *H. roshevitzii*, accession no. 7039, are as described previously (Kankaanpää *et al.* 1996). Genomic DNA was isolated, spiked with bacteriophage lambda DNA, and blots were prepared as earlier (Vicient *et al.* 1999). Four replicates of 20, 10, 5, and 2.5 ng were blotted for each accession. Blots contained calibration controls consisting of LARD LTR amplified with primers 91673 and 91674 described above. The PCR fragments were loaded in three replicates each of 100, 50, 10, and 1 pg, with 10 ng herring sperm DNA added as carrier. Samples were cross-linked to the filters following blotting. Probes for the LARD LTRs consisted of the same PCR product that was blotted and were ^{32}P labeled with Rediprime II (Amersham) as described above. Prehybridizations were carried out for 2 hr at 65° in 5× SSC, 0.5% SDS, 5× Denhardt's reagent, and 20 ng ml^{-1} herring sperm DNA. Hybridizations were carried out overnight at 65° in the same buffer with the addition of 10 ng radiolabeled probe. Filters were washed twice with 2× SSC, 0.1% SDS at room temperature for 10 min, followed by twice in 2× SSC, 0.1% SDS at 65° for 10 min, and finally in 0.1× SSC at 65° for 20 min. Bound radiation was quantified on a phosphoimager as described before (Vicient *et al.* 1999). Hybridized probe was removed by washing the filters in 0.2 M NaOH, 0.1% SDS twice at 37° for 20 min to prepare blots for rehybridization. The amount of bound DNA for each sample was normalized by hybridization to a lambda DNA probe, and the response to each probe was calculated as described earlier (Vicient *et al.* 1999).

Relative LTR and internal domain abundance: Amplification reactions were carried out between the LTR and (−)-strand priming site (PBS) using the primers LTROL2, above, and primer InTol, 5′-GGTTCGATCAATCAAGGGGGCTC-3′. The LTR fragment was produced with the primers Gossy-L, 5′-ATATCTTGTGCATCGGGATTCC-3′, and Gossy-R, 5′-GAC ATAACCCCACCGTGTCCTC-3′. Amplification was carried out in a 20-μl reaction containing 75 mM Tris-HCl, pH 8.8, 20 mM $(NH_4)_2SO_4$, 1.5 mM $MgCl_2$, 0.01% Tween-20, 20 ng barley DNA, 200 nM each primer, 200 μM dNTPs, and 1 unit Taq DNA polymerase. The amplification program consisted of 94°, 2 min; 7–16 cycles of 94°, 20 sec, 56°, 20 sec, 68°, 30 sec; and a final elongation of 68°, 5 min. One-fifth of each reaction (4 μl) was separated by electrophoresis in 2% agarose gels (RESolute wide range agarose; BIOzym, Landgraaf, The Netherlands) by electrophoresis (80–100 V) and visualized by ethidium bromide.

Cloning of LTR flanks: Ligation-mediated PCR (Siebert *et al.* 1995) was carried out on selected genomic DNAs obtained as above, using the GenomeWalker kit (Clontech Laboratories, Palo Alto, CA) according to the manufacturer's instructions.

Fluorescent *in situ* hybridization: Chromosomes were prepared from barley root tips as previously described (Vicient *et al.* 1999), and *in situ* hybridizations were carried out as before (Anamthawat-Jónsson *et al.* 1996). Chromosomes were denatured in 70% formamide, 2× SSC for 3 min at 74°. Preparations were probed with the LTR segment of the *Sukkula* LARD element, described for the BAC hybridizations above. The probes were labeled with biotin, denatured in 40% formamide, and boiled for 5 min. Hybridization was carried out at 37° overnight at a stringency of 77% (high), and hybridized probe was detected with ExtrAvidin-FITC (Sigma, St. Louis) as described earlier (Vicient *et al.* 1999).

Data mining, alignment, and sequence analysis: Preliminary sequence alignments were in some cases made with the PILEUP program available in the Wisconsin package version 10.2 (Oxford Molecular Group) available on the irix server (Center for Scientific Computing, Espoo, Finland; http://www.csc.fi/suomi/info/index.phtml.en). Sequences were aligned using the CLUSTALW multiple alignment program (Thompson *et al.* 1994; Higgins *et al.* 1996), either online at http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_clustalwan.html or locally, using the ClustalX version 1.8 downloaded from http://www-igbmc.u-strasbg.fr/BioInfo/. The program was also used to search for overall variability. Final alignment editing for display was carried out with the GeneDoc version 2.6 (K. B. Nicholas and H. B. Nicholas, http://www.psc.edu/biomed/genedoc). DNA sequences were analyzed with the BLASTN and BLASTX algorithms against public DNA and protein sequence databases, online at the National Center for Biotechnology Information (NCBI) via http://www.ncbi.nlm.nih.gov/BLAST/. The GENSCAN program was used for more refined searches on sequences that had been identified by BLASTN and BLASTX as having homology to other genes in the nucleic acid database, using the Massachusetts Institute of Technology server located at http://genes.mit.edu/GENSCAN.html.

DNA and RNA information content: To find the relevant signal in the multiple alignments we used the Kullback-Leibler measure (Kullback and Leibler 1951; Cover and Thomas 1991). The measure quantifies the contrast between an actual and an expected distribution of nucleotides. This is used to calculate the total amount of information per position in the alignment. Neglecting gaps and using background distribution in this measure reduced to the Shannon information measure used by Schneider and Stephens (1990). For gaps, we used the background probability $p = 1$ as others have (Kullback and Leibler 1951; Schneider and Stephens 1990; Cover and Thomas 1991). For the Shannon information measure, the maximum information in bits per position is $\log_2 4 = 2$ for nucleotide sequences. The quantifier p_k used here for nucleotides is $p_n = 0.25$.

Analysis of multiple alignments of RNAfold predictions: Candidate regions of RNA secondary structure were provided by the Mfold, version 3.0 (Zuker and Stiegler 1981; Zuker 1989, 1994), and RNAfold, Vienna RNA Package version 1.4. (Jaeger *et al.* 1990; Hofacker *et al.* 1994). Conservation of an RNA secondary structure element in position i is calculated again using a relative information measure,

$$I_i = \sum_{k=-,.,(,)} q_{ik} \log_2 \frac{q_{ik}}{p_k}, \qquad (1)$$

where the index k runs over all RNA secondary-structure elements and gaps. The quantities $q_{i,(}$, $q_{i,)}$, $q_{i,.}$, and $q_{i,-}$, are the observed fractions of 5 double strands (ds), 3 ds, single

strands (ss), and gaps correspondingly at position *i*. The expected probability of single-stranded RNA at every position *i*, $p_.$, has been found empirically to equal 0.5; consequently ds probabilities $p_(, p_)$ are equal to 0.25, and the gap background, $p_–$, is equal to 1.

Visualization of RNAfolds: To illustrate the most conserved features of the putative RNA secondary structure in the region between positions 200 and 400 in the multiple alignment file of LARDs we made two types of RNAfold predictions: a common RNA secondary structure and the specific representative of each fragment. The putative common RNAfold was predicted by the GeneBee (Brodsky *et al.* 1995), a server for RNA secondary structure prediction based on sequence alignment. The aligned restricted fragments were extracted from the aligned LARD sequences described above. To illustrate the folding of each sequence, the Mfold server was used on the region of interest extracted from each sequence.

RESULTS

LARD LTRs have long, conserved ends but highly variable centers: Primers were designed to match the ends of the LARD LTR sequences available in the database. These primers produced bands of 4.1–4.5 kb from template DNA of all Hordeum species investigated (data not shown). A set of species from the tribe Triticeae of the Poaceae was chosen for examination. The Triticeae (Barkworth 1992; Hsiao *et al.* 1995; Petersen and Seberg 1997) include the genera of several of the most important cereals, barley (*Hordeum vulgare* L.), wheat (*Triticum sativum* L.), and rye (*Secale cereale* L.), and therefore include the closest relatives to the source of the original *Sukkula* element. The products were cloned and sequenced from barley, *H. marinum*, *H. murinum*, *H. patagonicum*, and wheat. The sequences, together with the similar ones in the database, ranged from 3130 to 5605 bp, yielding an average length of 4484 ± 687 bp [standard error of the mean (SEM)]. This is unusually long for LTRs. Because the central regions of these sequences are highly variable, alignments were constructed for the 5′ and 3′ ends, oriented with respect to the predicted direction of transcription.

The alignment of the 5′ end of the LTR, consisting of 2734 nucleotides (nt; accession ALIGN_000282; access via ftp://ftp.ebi.ac.uk/pub/databases/embl/align/) reveals a highly conserved segment extending from the 5′ universal terminus, 5′-TGTGACAGCCCGA . . . . -3′, for ~400 nt into the element (Figure 2A). The region is conserved also in *H. marinum* and *H. murinum*, which belong to different genome groups than do barley and *H. patagonicum* and are not closely related to them. The sequence available for a similar element, tasuk10 from wheat (AF029897), is quite divergent from the others and may be degenerate. Alignment of the succeeding segments of the sequences reveals a highly conserved stretch extending inward for ~1880 nt from the 3′ end of the LTR (Figure 2B and accession ALIGN_000280). Hence, ~2250 bp, or ~50%, of the ~4.5-kb LTR of LARD elements is conserved, but the region that is variable in the alignment is variable even among the sequences determined from the barley genome alone.

Retrotransposon LTRs contain short inverted repeats at their termini. Because the PCR primers were designed to match the termini, the actual termini in the sequenced products shown in Figure 2 cannot be analyzed. However, the large clones containing LARD LTRs present in the database, as well as additional sequence data we have collected using subterminal LTR primers, define imperfect 6-bp terminal inverted repeats: 5′-TGT RAC . . . . . GTRACA-3′. These database accessions also universally display the conserved terminal 5′-TG . . . . . CA-3′, found in all retroviruses and LTR-containing retrotransposons, and contain 5-bp direct repeats flanking the termini.

LARD internal domains are highly conserved but noncoding: To amplify the internal region lying between the left and right LTRs, primers were designed so that the 5′ primer matched the 3′ end of the LTR and the 3′ primer matched the 5′ end of the LTR. Amplifications were carried out on genomic DNA from 13 species in four genera in the tribe Triticeae, including barley, and generated single bands of ~3.5 kb. In all 35 sequences aligned (alignment database accession ALIGN_000601), a highly conserved domain matching the initiator-methionyl tRNA ($^{i}_{met}$tRNA) was found immediately prior to the 5′ LTR (Figure 3A). A domain in this position homologous to tRNA is almost universal among retrotransposons and retroviruses, where it serves as the minus-strand priming site (PBS) for synthesis of the initial cDNA strand during reverse transcription (Marquet *et al.* 1995). The $^{i}_{met}$tRNA is one of the most commonly used. The consensus for LARDs is 5′-TGGTATCA GAGCC-3′. The first 11 nucleotides and the thirteenth are completely conserved between the sequenced elements and $^{i}_{met}$tRNA. The twelfth, a cytosine, is missing in four sequences, those from barley. The same nucleotide was highly variable in the PBS sequences of 11 retrotransposon families compared earlier (Suoniemi *et al.* 1997).

Immediately internal to the 3′ LTR, functional retroelements contain a conserved segment rich in guanine and adenine referred to as a polypurine tract (PPT). The PPT serves as the priming site for the reverse transcription of the plus-strand, the second strand of the cDNA to be synthesized (Friant *et al.* 1996; Schultz *et al.* 2000). In an alignment of this part of the LARD internal domains (Figure 3B), the PPT is highly conserved, composed of 11 purines in a 12-nucleotide stretch, 5′-AGTGGAGGAGAG-3′. The consensus for the end of this segment is 5′-GGGAG-3′ in a broad range of retrotransposons and retroviruses (Suoniemi *et al.* 1997) and is the form found in the oat sequences examined here. This is much different from the 5′-GAGG GGGCG-3′ reported for TRIMs (Witte *et al.* 2001). The well-conserved PPT and PBS domains suggest that

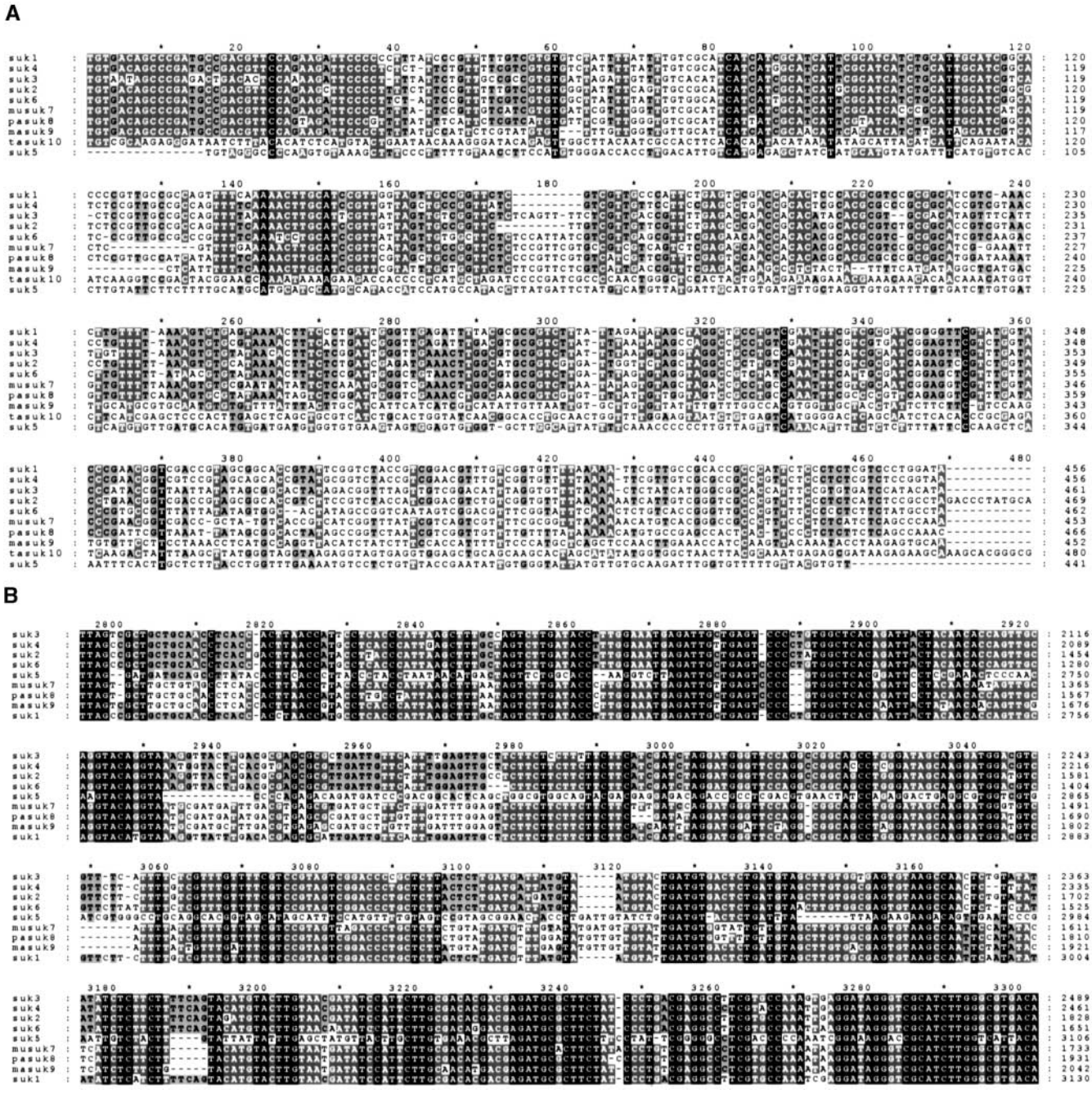

FIGURE 2.—Sequence alignment of the terminal 5′ (A) and 3′ (B) regions of LARD LTRs. The alignment includes sequences from barley (suk1, suk2, suk3, suk4, suk5, and suk6), *Hordeum murinum* (musuk7), *H. patagonicum* (pasuk8), *H. marinum* (masuk9), and *Triticum aestivum* (tasuk10). Shading represents the degree of sequence conservation at each nucleotide position: white letters on black, present in 100% of the sequences; white on gray, in 80–100%; black on gray, in 60–80%; black on white, in <60%. Complete alignments, including constituent sequence information, are available at ftp://ftp.ebi.ac.uk/pub/databases/embl/align/ for A as accession ALIGN_000282 and for B as ALIGN_000280.

LARD elements retain the capacity for being reverse transcribed.

The entire internal domain was determined for these 35 sequences. Beyond the PBS, the internal domain was highly conserved on the DNA level for almost all species in a block extending to position 710 in the alignment (alignment accession ALIGN_000601). All five oat sequences lacked a segment from nucleotides 100 to 1010. The Triticum sequences and those of *Elymus repens*, *H. roshevitzii*, *H. marinum*, and one barley sequence lacked the block extending from nucleotide 740 to nucleotide 920 in the alignment. All sequences except those of oat were well conserved from nucleotide 920 to 1140. The sequences of all accessions from nucleotides 1160 to

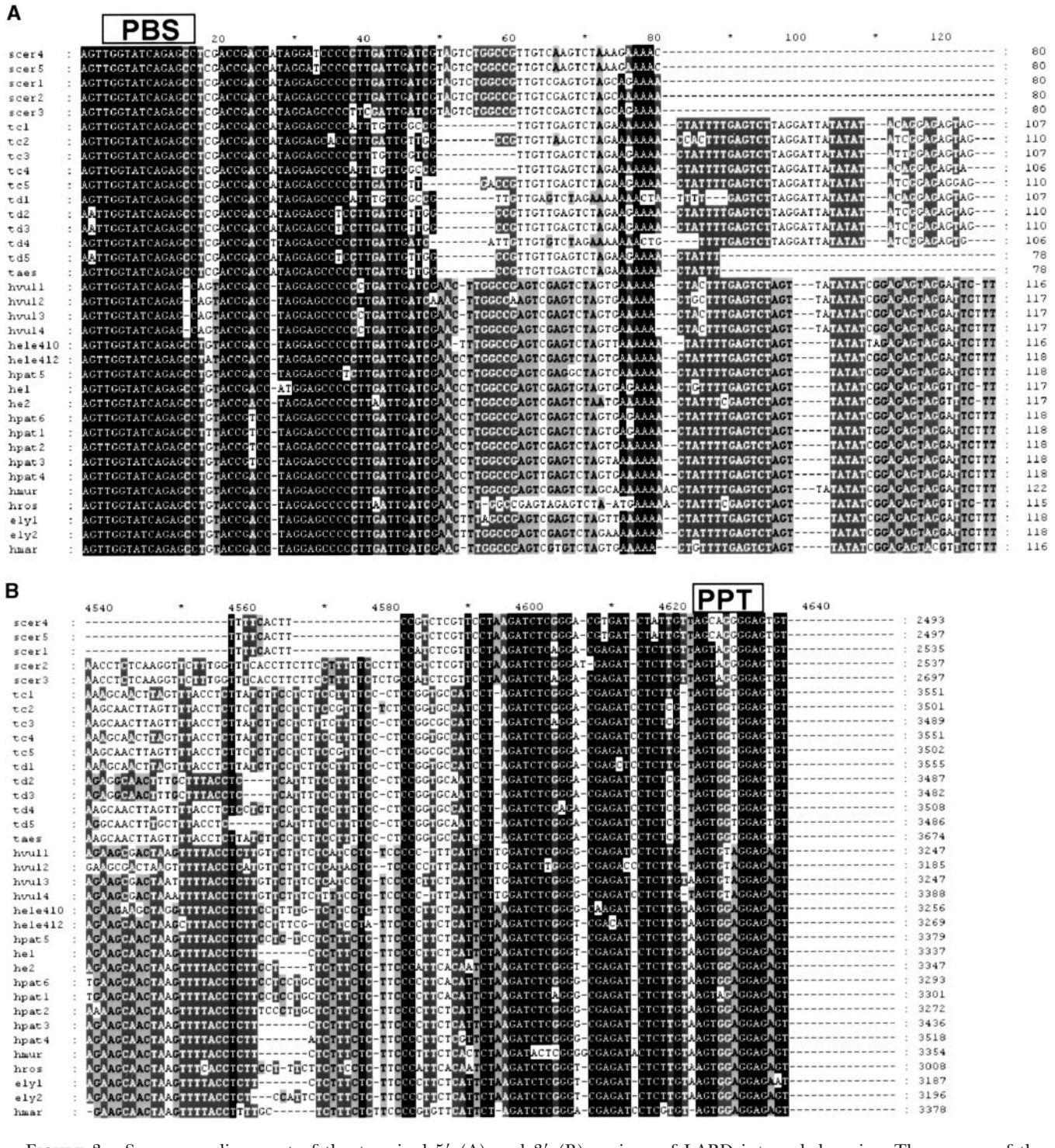

FIGURE 3.—Sequence alignment of the terminal 5′ (A) and 3′ (B) regions of LARD internal domains. The source of the aligned sequence is encoded within the prefix of the name: scer, *Secale cereale* (rye); tc, *Triticum dicoccoides*; td, *T. durum* (durum wheat); taes, *T. aestivum* (bread wheat); hvul, *Hordeum vulgare* (barley); hele, *H. erectifolium*; hpat, *H. patigonicum*; hmur, *H. murinum*; hros, *H. roshevitzii*; ely, *Elymus repens*; and hmar, *H. marinum*. The positions of the (−)-strand priming site, called the PBS, and of the (+)-strand priming site, called the PPT for polypurine tract, are shown above the corresponding sequences in A and B, respectively. Shading represents the degree of sequence conservation at each nucleotide position: white letters on black, present in >90% of the sequences; white on gray, in 90–70%; black on gray, in 70–50%; black on white, in <50%. A complete alignment, including constituent sequence information, is available at ftp://ftp.ebi.ac.uk/pub/databases/embl/align/ for internal domain as ALIGN_000601.

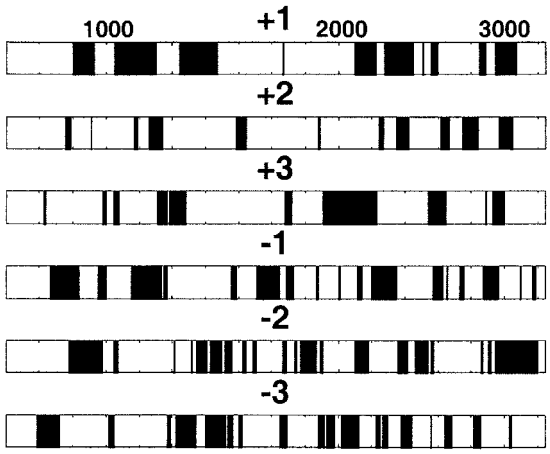

FIGURE 4.—Open reading frames in LARD internal domains. The potential coding regions in the LARD internal domain of barley hvul1 (accession AY054377) are displayed as solid boxes in each forward (+) and reverse (−) reading frame. Translationally competent *copia*- or *gypsy*-like retrotransposons generally have one or two long open reading frames (see Figure 1) in the forward orientation only.

2320 are well conserved, with insertions or deletions in individual sequences or species. From nucleotides 2340 to 2660 is poorly conserved in all sequences, but highly conserved between 2660 and 2880 and then poorly conserved until nucleotide 3200 in the alignment. The regions from 3500 to 3920 and the terminal region from 4300 to 4418 are highly conserved.

Given a fairly high degree of conservation between elements in different genera, we expected that the internal domains of the LARD elements would contain open reading frames (ORFs) coding for the products universally found in LTR retrotransposons, including integrase and reverse transcriptase. However, searches for ORFs in all six frames failed to reveal any that are sufficiently long to encode the major retrotransposon proteins, which are furthermore generally expressed as a polyprotein. We present one example from barley, the best characterized for retrotransposons of the species considered, typical of the other sequences (Figure 4). In the sense orientation with respect to the reverse transcriptase priming sites, 18 ORFs were longer than 50 nt, the longest 4 being 330, 255, 234, and 174 nt, respectively. Although it is conceivable, as seen in Figure 4, that coding domains for the expected proteins might be split among the reading frames by indels or stop codons, BLAST searches against the protein, DNA, and expressed sequence tag (EST) databases both for the internal domains and for the individual ORFs failed to yield any significant matches to retrotransposon or retroviral products. Therefore, the internal domains are simultaneously well conserved on the DNA level but noncoding. For this reason, the group has been named LARDs.

LARD elements are transcribed: The *Sukkula* LARD LTR matched the greatest number of ESTs of any retrotransposon query sequence examined (VICIENT *et al.* 2001a). The *H. vulgare* and *T. aestivum* database libraries, in which these ESTs were present, together contained roughly similar numbers of sequences as the Zea and Oryza libraries individually, yet none of the characterized maize or rice retrotransposons were as prevalent as LARDs among the ESTs. This suggests that LARD LTRs are efficient promoters. More recent BLAST searches using LARD query sequences from *H. vulgare*, *H. murinum*, *H. patagonicum*, and *H. marinum* (data not shown) identify ESTs matching both the LTR and internal domains of the elements.

LARD internal domains form well-conserved RNA secondary structures: The conservation on the sequence level of the internal domain, in the absence of protein-coding capacity, led us to explore other features possibly related to sequence conservation. Expansion and contraction of sets of repeat units, for example, might lead to homogenization of a region of DNA. However, when we analyzed the sequence for direct and inverted repeats using COMPARE and DOTPLOT of the Wisconsin package (data not shown), no repeats longer than 20 nt of either kind were found, even at a stringency of 14 matches in a scanning window of 21 nt.

The internal domains were then subjected to statistical analysis for information content. Information content is a convenient way to quantify the sequence conservation because it is additive when the positions are independent (SCHNEIDER *et al.* 1986; http://www.lecb.ncifcrf.gov/toms/) and can be used to approach problems in molecular evolution (ZEEBERG 2002). As applied, it is a logarithmic function of the decrease in uncertainty between a given state of conservation, for example, where only one base is found, and a fully unconserved or unselected sequence, where any of the four bases is found. As a logarithmic function and a comparison between two states, it can take on fractional values, as well as values <0 when gaps are considered.

Thirteen LARD internal regions were aligned by CLUSTALW and to get the total sequence conservation the information content was summed across all positions in alignment (SCHNEIDER *et al.* 1986; SCHNEIDER 1999). The plot shows that the LARD internal domains are divided into regions of respectively higher and lower information content, both for the DNA sequences, where it is correlated with sequence conservation (Figure 5A), and for RNA secondary structure. The information content, or sequence conservation, on the DNA level is high for most of the internal domain, except for two small regions around nucleotides 1800 and 2150. These regions also have low information content for the RNA (see below). In contrast, the information content of the conserved parts of the central region is comparable to that of the 5′ and 3′ termini that respectively contain the PBS and PPT, both of which are highly conserved (Figure 3). The regions are correlated, in the

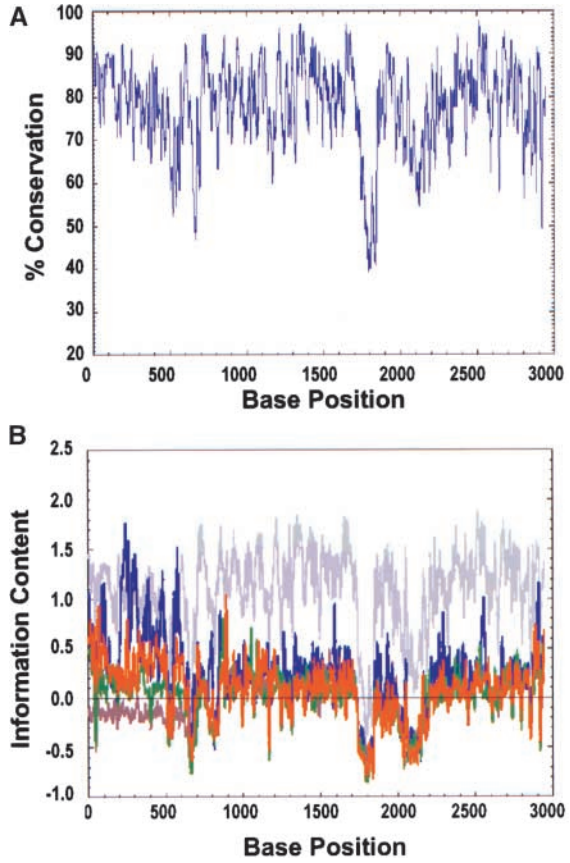

FIGURE 5.—Conservation of information content in the DNA sequences and RNA secondary structure of LARD internal domains. Sequences from *Eleymus repens* (sequence ely2, accession no. AF453662), *Hordeum vulgare* (hvul1, AY054377; hvul2, AY054381), *H. erectifolium* (hele1, AY069966), *H. euclaston* (he1, AF453660), *H. murinum* (hmur, AY054379), *H. patagonicum* (hpat1, AY054380), *H. roshevitzii* (hros, AF453670), *Triticum aestivum* (taes, AF453674), *T. dicoccoides* (tc1, AF453675), and *T. durum* (td1, AF453683) were aligned. (A) Plot of sequence conservation as percentage of match by nucleotide position over the internal domain. (B) Plots of information content. Sequence conservation was calculated as the information contribution of each base relative to its expected distribution according to Equation 1 (see text). The information content, a logarithmic function, for the DNA sequence (gray line) was smoothed by a running average with a window size of 6 nt. RNA secondary structure was predicted for each of the DNA sequences with the Vienna Package (HOFACKER *et al.* 1994), the outputs were aligned, and then gaps were inserted according to the DNA multiple alignments. The conservation of RNA secondary structure (blue line) at every position was computed as the information contribution of stem or loop relative to its expected distribution (see text). The curve showing the information content was smoothed by a running average with a window size of 6 nt. The procedure was repeated for RNA structures predicted for randomized sequences shuffled along the length of each sequence (green line) and shuffled at each position between sequences (orange line).

overlaid plot of Figure 5B, with regions of conserved RNA secondary structure.

The information content, or nonrandomness, of the RNA structure was compared with that for RNA corresponding to shuffled sequences (Figure 5B). Sequences were individually randomized (Figure 5B, "shuffled DNA") or randomized by swapping between aligned sequences at each individual position (Figure 5B, "vertically shuffled DNA"). In much of the internal domain, the predicted secondary structure is considerably above that for randomized sequence, particularly between nucleotides 1 and 500, 1200 and 1700, 1900 and 2100, and 2300 and 2500. Furthermore, of the regions with highest information content in the RNA secondary structure, respectively, at nucleotides 200, 500, 1500, 2200, 2500, and 2900 (Figure 5A), only those at nucleotides 2500 and 2900 have DNA conservation in the alignment (Figure 5B) >90%, the others being <70%. The region of high information content in the RNA structure at nucleotide 500 has DNA conservation of only 55%.

We concentrated further analysis on the region of low variability and high information content in the 5′ end of the internal domain and searched for candidate regions forming secondary structure with the programs Mfold and RNAfold. Multiple alignments were made of the RNA secondary structures to reveal RNA motifs common for the various LARD sequences in a way similar to our approach earlier (PELEG *et al.* 2002). In that study, we revealed a novel RNA secondary structure in the C1 region of human immunodeficiency virus (HIV)-1. The structural analyses performed on the 3′ regions (Figure 6) of the LARDs predicted a large hairpin of 200 bp with six main loops in most sequences and every energy level. Multiple alignments were made of the RNA secondary structures to reveal RNA motifs common for the various LARD sequences. These results were confirmed by using "Alifold" software, a method for computing the consensus structure of a set of aligned RNA sequences (HOFACKER *et al.* 2002). It takes into account both thermodynamic stability and sequence covariation. This software predicted an RNA secondary structure in positions 227–376. However, it detected more than one RNA secondary structure between nucleotides 227 and 505. This finding is reinforced by the information content plot of RNA secondary structure.

Genome prevalence and organization of LARD elements in barley: The copy number for LARDs in the barley genome was estimated by dot blot as before for *BARE*-1 (VICIENT *et al.* 1999). To account for any sequence variability, the hybridization probes consisted of internal and LTR fragments amplified from the genome with conserved primers. Variation in DNA loading or blotting was normalized through the addition of lambda DNA to plant genomic DNA preparations. The haploid genome of barley cv. Bomi (4.53 pg; VICIENT *et al.* 1999) is estimated to contain $1.42 \pm 0.11 \times 10^4$ (SEM) copies of the LTR, about the same as the number (1.32×10^4) of full-length *BARE*-1 elements in the barley genome but 6-fold less than the 9.0×10^4 *BARE*-1 LTRs (VICIENT *et al.* 1999). The wild species *H. roshevitzii* contains $2.06 \pm 0.18 \times 10^3$ (SEM) LTRs per haploid genome equivalent,

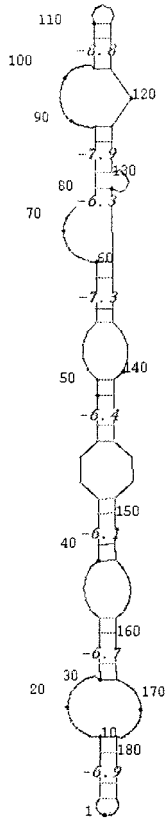

Figure 6.—A common RNA secondary structure prediction for the 5′ region of LARD internal domains. Each RNA sequence from position 200 to 400 in the multiple alignment was degapped, and the segment was folded. A large hairpin, 200 bp long, with six main loops along the structure, is the main feature of this putative common RNAfold. The free energy of the structure equals −19.0 kkal/mol.

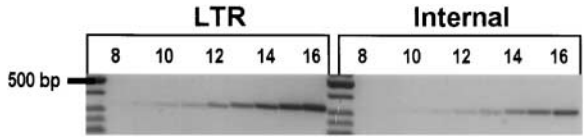

Figure 7.—Relative abundance of LARD LTRs and internal domains by PCR amplification. Fragments from the LTR (left) or extending from the LTR into the PBS region (right) of nearly identical size were amplified by a successively increasing number of cycles (displayed from left to right). The reactions generated a single product. A negative fluorescence image of a gel stained with ethidium bromide is shown.

8-fold less than the number of *BARE*-1 elements and 175-fold less than the total number of *BARE*-1 LTRs.

We have estimated the relative abundance of LARD LTRs and internal domains in barley cv. Bomi by PCR. Two sets of primers were designed, one to amplify LTR fragments and the other, matched in T_m and product size, to amplify a fragment bridging the LTR and the PBS region. The amplification efficiency of the two sets of primers was tested with equimolar amounts of cloned template and the results were used in normalizing the data. Amplification efficiency was assessed by densitometry at varying numbers of cycles (7–16), taking the logarithmic portion for analysis. The primer pair detecting the internal region required 1.84 cycles more for the same intensity of product fluorescence as the LTR, and the internal domain controls amplified 1.6 times more efficiently than the LTR (Figure 7). This indicates that the LARD LTRs are $2^{1.6+1.84}$ or 10.9 times more abundant in the genome than are internal domains. Because intact LTR retrotransposons contain two LTRs flanking an internal domain, a genome containing few solo LTRs of a given family of retroelements should display a ratio of LTRs to internal domains near 2:1. Hence, the haploid barley genome appears to contain only ~1300 full-length LARDs and 11,600 solo LTRs. This yields a solo LTR:full-length ratio of 8.9:1 for LARDs in barley. Previous studies have indicated that solo LTRs, produced by LTR-LTR recombination, are relatively common in the barley genome, particularly for retrotransposon *BARE*-1 (Vicient *et al.* 1999; Kalendar *et al.* 2000; Shirasu *et al.* 2000). Cultivated barley can have a ratio >11:1 for *BARE*-1, indicating an abundance of solo LTRs for that family of elements as well (Vicient *et al.* 1999).

On the local scale, both the original *BARE*-1 insertion element (Manninen and Schulman 1993) and the two *Sukkula* elements (Shirasu *et al.* 2000) in a 66-kb contiguous sequence in barley (AF254799) are solo LARD LTRs nested in other retroelements. Of the other LARD LTRs in the database, the solo *Sukkula* on a 124-kb BAC (AF474373) is flanked by unknown sequence and that on a 60-kb stretch in the *Mlo* region of chromosome 4H (Y14573) matches *Sabrina* (Shirasu *et al.* 2000) and *Wham* retrotransposon sequences from a *T. monococcum* BAC (AF459639) and may represent a retroelement nest. To extend this data set, we carried out adapter-mediated PCR to clone and sequence LTR flanks (Siebert *et al.* 1995), using adapters corresponding to different restriction enzymes. Of the 12 sets of flanks we obtained, 10 represent insertions into other LARD LTRs, one is an insertion into a *Bagy2* retrotransposon (AJ279072), and one set of flanks is an unknown sequence (data not shown). These data indicate that LARD elements behave like other retrotransposons in barley (Shirasu *et al.* 2000) and elsewhere (SanMiguel *et al.* 1996; Wicker *et al.* 2001), forming nests of inserted elements.

Insertions nested into retrotransposons and other repeated DNA make it difficult to prove that the LARD polymorphisms (Manninen *et al.* 2000; Boyko *et al.* 2002) are due to new insertions, because the empty sites are not unique. However, of five sequenced LARD insertions in the database (AF47473, AF254799, AF474072, and Z71327), all contain conserved direct repeats of the sort generated by integrase activity (data not shown). These insertions were not chosen for polymorphism (selecting for recency) but happen to be in database clones for gene-rich regions (where insertions of high-

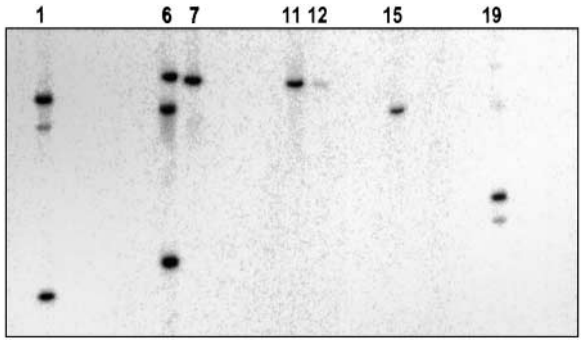

FIGURE 8.—DNA gel blot analysis of 20 barley genomic BAC clones. Clones were digested with *Xba*I and the blots were probed with a *Sukkula* LARD LTR fragment.

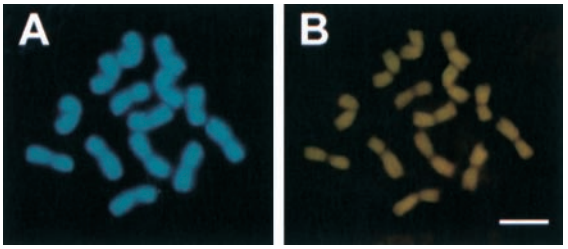

FIGURE 9.—Fluorescent *in situ* hybridization of a *Sukkula* LARD LTR probe to barley chromosomes. (A) 4′,6-Diamidino-2-phenylindole staining of total DNA. (B) FITC fluorescence on the chromosomes after hybridization with the labeled LTR probe. Bar, 5 μm.

copy-number retrotransposons are not favored). This provides support for the recent activity of LARD elements in barley.

Using a LARD LTR probe, 7 of 20 BAC clones gave a total of 12 hybridizing bands (Figure 8). For comparison, 19 of the same 20 BAC clones contained a total of 50 *BARE*-1 LTRs, and 9 of 20 contained internal domains (VICIENT *et al.* 1999). The disparity suggests that *Sukkula* LTRs, in addition to being rarer than *BARE*-1, are not as evenly distributed in the genome. Fluorescent *in situ* hybridization to barley chromosomes, made with a *Sukkula* probe (Figure 9), also indicates a high copy number for the LTRs. The LTR probe uniformly labels the chromosome arms except at the telomeres, nucleolar organizing regions, and centromeres, where the signal is lacking. This pattern is similar to that seen earlier for *BARE*-1 (SUONIEMI *et al.* 1996; VICIENT *et al.* 1999), but contrasts sharply with retroelements such as Cereba (HUDAKOVA *et al.* 2001) that are localized to the centromere or paracentric regions. The chromosome hybridization data were confirmed by PCR amplifications on genomic DNA isolated from wheat-barley chromosome addition lines (data not shown). We used PCR primers specific for both LTRs and internal domains of barley and analyzed segments of the barley genome present in a wheat (*T. aestivum*) cv. Chinese Spring genomic background. Both the LTR and internal domains were present on all barley chromosomal segments.

DISCUSSION

LARDs, a distinct group of retroelements: The LTR retrotransposons and retroviruses generally display conservation of structure against a background of fairly rapid sequence evolution. The RNA processing signals, encoded proteins, and arrangement of the coding domains are remarkably conserved among the major groups of these elements in the eukaryotes (XIONG and EICKBUSH 1990; SUONIEMI *et al.* 1998a; KUMAR and BENNETZEN 1999; VICIENT *et al.* 2001b). These features appear to have been preserved since before the divergence of the evolutionary lines leading to the modern plants, animals, and fungi. However, individual retroelements are frequently translationally or functionally incompetent due to the accumulation of mutations (SUONIEMI *et al.* 1998b). Such elements may be mobile nevertheless if they are complemented *in trans* by the retrotransposon proteins needed for reverse transcription, packaging, and integration.

A group of small, nonautonomous retroelements, TRIMs, was recently reported (WITTE *et al.* 2001). These contain very short LTRs or LTR deletion products, of 100–250 bp and internal domains of 100–300 bp comprising the priming sites for reverse transcription and little else. We have described here a new group, LARDs. The LARDs, like the TRIMs, lack protein-coding capacity and are therefore nonautonomous. However, TRIM elements lack an internal domain virtually completely, while LARDs have large, conserved, noncoding, internal domains of 3.5 kb. Furthermore, while TRIMS have small LTRs, LARDs have LTRs of ~4.5 kb, among the longest known. Each group is conserved, and elements intermediate between TRIMs and LARDs have not been found. Therefore, the elements in these two groups appear to follow different replicational or life-cycle strategies and probably have distinct histories. The recently reported Dasheng element of rice is of the LARD type (JIANG *et al.* 2002a,b).

LARD internal domains produce conserved RNA structures: A remarkable feature of the LARD element sequences, determined for 13 species, is the conservation of the noncoding central domain. Folding predictions for the internal domain revealed a highly conserved common RNAfold in 12 out of 13 species. The predicted RNA transcription products for the elements sequenced all show a hairpin loop of ~200 bp starting 200 bp downstream from the 3′ end of the 5′ LTR. The observed conservation of the secondary structure in the LARD internal regions could not be easily explained as an outcome of high sequence conservation; a rather small number of differences in homologous sequences can cause dramatic changes in secondary structure (FONTANA *et al* 1993). Furthermore (BONHOEFFER *et al.*

1993; Fontana *et al.* 1993), even a 10% sequence difference may yield secondary structures that display, at most, obscure similarities. This has been demonstrated here for the LARDs (Figure 5B), where we compare LARD RNA secondary structure conservation with that for randomized sequences. Not only do randomized LARDs show no common structures, but also "vertical" shuffling of the aligned LARDs completely annihilates common secondary structures. The procedure of vertical shuffling of aligned sequences preserves both a total number of mutations and the value of information content in every aligned position but this positional sequence conservation does not extend to RNAfold conservation.

Thus, positional conservation of a stem-loop structure among related sequences should be seen as a consequence of the biological function of the RNA structure rather than as a consequence of sequence homology. The high conservation of this 5′ hairpin in LARD internal domains suggests a biological function, which is still unknown. Hairpin structures have been predicted for the LTRs of plant (Lucas *et al.* 1992) and other retrotransposons. The *Grande1* maize retrotransposons contain long repeats with conserved stem-loop structures (Monfort *et al.* 1995). Stem-loop structures or hairpins in transcripts have been implicated in a variety of roles in retroviruses such as HIV, where small hairpins serve as packaging signals (Kerwood *et al.* 2001), in reverse transcription (Driscoll *et al.* 2001) and transcription (Roebuck and Saifuddin 1999), and in nuclear export of RNA (Yang *et al.* 2000). In general, the conservation of DNA coding capacity and the conservation of RNA secondary structures do not tend to demonstrate the same pattern in parallel. This is not the case with the LARDs, where the DNA conservation pattern follows that of the RNA secondary structure with a high correlation coefficient. These considerations suggest that the LARD sequence conservation is related to the structural conservation, which in turn is connected to its function as RNA. Analyses of the conserved RNA structures in the rest of the internal domain are in progress and will be reported separately.

LARD elements are abundant in the barley genome: A combination of DNA gel blot hybridizations, BAC clone analyses, PCR amplification experiments, and fluorescent *in situ* hybridizations indicate that LARD elements are abundant in the barley genome. We estimate that there are 1.3×10^3 full-length LARDs, and 1.2×10^4 solo LARD LTRs such as *Sukkula*, in the haploid barley genome. Full-length *BARE*-1 retrotransposons, by contrast, are present in $\sim 1.32 \times 10^4$ copies (Vicient *et al.* 1999). Recombination generates solo LTRs, but their frequency may vary greatly from retrotransposon to retrotransposon. In maize, the *Huck* element family appears to contain many solo LTRs (Meyers *et al.* 2001), but excepting *Ji* others may not (SanMiguel *et al.* 1996).

It was observed previously (Shirasu *et al.* 2000) that the propensity of LTRs to undergo recombination may vary directly with their length. The LARD LTRs are very long, ~4.5 kb, although only 400 bp at the 5′ end and 1900 bp at the 3′ end are well conserved between elements. The conserved 3′ end is about as long as *BARE*-1 LTRs are in total. The ratio of solo LTRs to full-length elements is ~11:1 for *BARE*-1 and 8.9:1 for LARDs, despite the difference in their copy number. This is consistent both with dependence on length and with the probability of recombination being independent for each element. Nested retrotransposons appear characteristic of barley and other cereals (SanMiguel *et al.* 1996; Shirasu *et al.* 2000; Wicker *et al.* 2001). Our data and those in the DNA databases indicate that LARD elements are not exceptions to this organization.

LARD elements are active: Not only are LARD elements abundant in barley, but also they are found throughout the Triticeae, the tribe that includes barley, and in grass species outside the Triticeae. In the Triticeae, in both barley and *A. tauschii*, the D-genome ancestor of bread wheat, the LARD elements are polymorphic in their insertion sites and serve well in recombinational map construction (Manninen *et al.* 2000; Boyko *et al.* 2002). In a barley cross, 15 out of 35 scored *Sukkula* bands, generated by the interretrotransposon-amplified polymorphism method (Kalendar *et al.* 1999), were polymorphic and could be mapped (Manninen *et al.* 2000). A total of 31 markers generated with a primer designed to a barley *Sukkula* sequence were mapped on the *A. tauschii* genome (Boyko *et al.* 2002). The LARD elements, which the EST databases indicate are transcribed, display at least as much polymorphism between barley varieties as does the transcriptionally and translationally active *BARE*-1 family. Of the four database accessions from barley for which LARD integration sites can be clearly identified, all five LARDs are flanked by identical direct repeats (data not shown). Although these insertions are not known to be polymorphic, conservation of the direct repeats indicates that they occurred relatively recently. These data are consistent with LARD elements having been integrationally active since genetic separation of the parents in the mapping crosses.

Which autonomous retrotransposon drives LARD mobility? The conserved features described above, combined with the polymorphisms in barley varieties for the LARD element *Sukkula* insertion sites, suggest that LARDs remain active despite their lack of coding capacity. The LARDs therefore appear to be very large nonautonomous retroelements. This interpretation raises the questions of which retroelement class LARDs belong to and which family may provide the complementary proteins required for LARD replication, packaging, and integration. One sequence with a good match by BLAST search criteria with the *Sukkula* LTR was found within a 211-kb contiguous sequence from *T. monococcum* (AF326781, nucleotides 10,4294–129,635). The matching region was described as *Erika*-1, a *gypsy*-like retro-

transposon (Wicker *et al.* 2001). The LTRs of *Erika*-1 have 69% identity with the 3′ two-thirds of *Sukkula* LTRs, but a poor match at the 5′ end. The *Erika*-1 internal domain is 63% identical on the DNA level to the *gypsy*-like *Bagy*-1 retrotransposon of barley (Y14573, nucleotides 44,637–59,060), to which it has been described as being similar (Wicker *et al.* 2001), and is in turn ~74% similar to the internal domain of barley LARD elements. Although *Erika* is interrupted by stop codons, the internal domain finds predicted or identified retroelement translation products in BLASTX searches against protein databases. *Bagy*-1 produces no extensive alignments with barley LARD elements and their LTRs are not similar. The *Bagy*-1 PBS, 5′-TGGTAACAGA-3′, differs by 1 nucleotide from the PBS of LARD elements. However, the *Bagy*-1 PPT, 5′-GAGGGGGTGAG-3′, corresponds poorly to that of the LARDs.

A "*Sukkula* polyprotein" has been annotated at the Mla locus (AF427791). The two *Sukkula* elements in the accession appear to be identical copies and partially deleted. Alignment on the DNA level of the region assigned at Mla to the Sukkula polyprotein to the internal domain of our reported LARDs gives only a 38%, and highly gapped, identity, using a Smith-Waterman method, and only a 50-bp region of 66% identity, using a best-local-alignment method. Hence, the relationship between the conserved internal domains of the LARDs reported here and the putative Sukkula polyprotein remains unclear.

Another *gypsy*-like element, *RIRE3* (AB014739) has been described as being similar to *Sukkula*, which the authors refer to as *BARE101* (Kumekawa *et al.* 1999). However, our alignment of the *RIRE3* LTR and *Sukkula* (data not shown) produces only a 278-bp region of 61% identity, not including the very dissimilar terminal sequences. Nevertheless, the *RIRE3* internal domain (AB014738) produces an alignment to the *Erika* internal domain on the DNA level of 2876 bp and 62% identity and an alignment of 1370 residues of 56% similarity on the protein level, making it likely that *Erika* is a *RIRE3*-like retrotransposon. Taking these data together, the strongest conclusion that can be reached is that LARDs are probably mobilized by a *gypsy*-like retrotransposon resembling *Erika*-1 of *T. monococcum* and *RIRE3* of rice, but the mobilizing element is unlikely to be *Bagy*-1.

Although we have described LARD elements primarily from barley and related grasses, research by Jiang and co-workers (Jiang *et al.* 2002a,b) indicates that LARDs are found in rice as well and are part of a two-element, autonomous and nonautonomous, system, with the *gypsy*-like *RIRE2* as the possible autonomous partner. The *RIRE2* element has no extensive homologies with Hordeum LARDs. In BLASTn searches (nonredundant search in GenBank release 136), the only significant match (E value 3×10^{-12}, 87% identity) is to a BAC clone (AY268139) in a stretch of 73 nucleotides annotated as part of the Lolaog_184G9-1p LTR retrotransposon. The Lolaog element is >95% identical to regions in at least six other BAC clones of barley; in one (AF427791) it is annotated as a partial *Bagy-2* element. Phylogenetic analyses of *gypsy*-like elements, however, place *Bagy-2* in a separate clade from *RIRE2* (Vicient *et al.* 2001b). Hence, on the basis of the rice LARDs it is difficult to establish the most likely autonomous partner for barley LARDs.

In conclusion, LARDs are a group of nonautonomous retroelements conserved in sequence and structure that appear to be part of a system involving complementation by a family of probably *gypsy*-like retrotransposons encoding the protein products necessary for replication and integration. It is interesting, in this context, that LARDs are abundant, raising the question of whether LARDs are parasitic on their corresponding autonomous elements or whether the transpositional success of the latter is unaffected by LARD propagation. The existence of two groups of nonautonomous retrotransposons, TRIMs and LARDs, differing significantly in size, raises interesting questions about their comparative replicational strategies and their role in autonomous-nonautonomous binary systems. To our knowledge, similar elements have not been found in animals or fungi. It remains to be seen if these groups of retroelements represent evolutionary innovations in the plants or if they have parallels elsewhere.

We thank Dr. Ole Seberg (Botanical Institute, Copenhagen, Denmark) and Boreal Plant Breeding Ltd. (Jokioinen, Finland) for gifts of plant materials. Andris Kleinhofs (Washington State University, Pullman, WA) is thanked for gifts of BAC clones. The expert technical assistance of Anne-Mari Narvanto is deeply appreciated. Prof. Eduard Trifonov is gratefully acknowledged for helpful discussions. This study was supported by the European Union Research Directorate (contract no. QLK5-CT-1999-01499) and by the Academy of Finland (project 44404).

LITERATURE CITED

Anamthawat-Jónsson, K., J. S. Heslop-Harrison and T. Schwarzacher, 1996 Genomic *in situ* hybridization for whole chromosome and genome analysis, pp. 1–24 in *In Situ Hybridization Laboratory Companion*, edited by M. Clark. Chapman & Hall, Weinheim, Germany.

Barkworth, M. E., 1992 Taxonomy of the Triticeae: a historical perspective. Hereditas **116:** 1–14.

Bonhoeffer, S., J. S. McCaskill, P. F. Stadler and P. Schuster, 1993 RNA multi-structure landscapes: a study based on temperature dependent partition functions. Eur. Biophys. J. **22:** 13–24.

Boyko, E., R. Kalendar, V. Korzun, B. Gill and A. H. Schulman, 2002 Combined mapping of *Aegilops tauschii* by retrotransposon, microsatellite, and gene markers. Plant Mol. Biol. **48:** 767–790.

Brodsky, L. I., V. V. Ivanov, Y. L. Kalaydzidis, A. M. Lenotovich, V. K. Nikolaev *et al.*, 1995 GeneBee-NET: internet-based server for analyzing biopolymers structure. Biochemistry **60:** 923–928.

Casacuberta, J. M., S. Vernhettes and M.-A. Grandbastien, 1995 Sequence variability within the tobacco retrotransposon *Tnt1* population. EMBO J. **14:** 2670–2678.

Cover, T. M., and J. A. Thomas, 1991 *Elements of Information Theory*. John Wiley & Sons, New York.

Driscoll, M. D., M. P. Golinelli and S. H. Hughes, 2001 In vitro analysis of human immunodeficiency virus type 1 minus-strand strong-stop DNA synthesis and genomic RNA processing. J. Virol. **75:** 672–686.

Flavell, A. J., E. Dunbar, R. Anderson, S. R. Pearce, R. Hartley *et al.*, 1992 *Ty1-copia* group retrotransposons are ubiquitous and heterogeneous in higher plants. Nucleic Acids Res. **20:** 3639–3644.

Fontana, W., D. A. Konings, P. F. Stadler and P. Schuster, 1993 Statistics of RNA secondary structures. Biopolymers **33:** 1389–1404.

Frankel, A. D., and J. A. Young, 1998 HIV-1: fifteen proteins and an RNA. Annu. Rev. Biochem. **67:** 1–25.

Friant, S., T. Heyman, F. X. Wilhelm and M. Wilhelm, 1996 Role of RNA primers in initiation of minus-strand and plus-strand DNA synthesis of the yeast retrotransposon Ty1. Biochemie **78:** 674–680.

Gabriel, A., and E. H. Mules, 1999 Fidelity of retrotransposon replication. Ann. NY Acad. Sci. **870:** 108–118.

Hartl, D. L., E. R. Lozovskaya and J. G. Lawrence, 1992 Nonautonomous transposable elements in prokaryotes and eukaryotes. Genetica **86:** 47–53.

Heidmann, T., O. Heidmann and J. F. Nicolas, 1988 An indicator gene to demonstrate intracellular transposition of defective retroviruses. Proc. Natl. Acad. Sci. USA **85:** 2219–2223.

Higgins, D. G., J. D. Thompson and T. J. Gibson, 1996 Using CLUSTAL for multiple sequence alignments. Methods Enzymol. **266:** 383–402.

Hofacker, I. L., W. Fontana, P. F. Stadler, L. Bonhoeffer, M. Tacker *et al.*, 1994 Fast folding and comparison of RNA secondary structures. Monatsh. Chem. **125:** 167–188.

Hofacker, I. L., M. Fekete and P. F. Stadler, 2002 Secondary structure prediction for aligned RNA sequences. J. Mol. Biol. **319:** 1059–1066.

Hsiao, C., N. J. Chatterton, K. H. Asay and K. B. Jensen, 1995 Phylogenetic relationships of the monogenomic species of the wheat tribe, Triticeae (Poaceae), inferred from nuclear rDNA (internal transcribed spacer) sequences. Genome **38:** 211–223.

Hudakova, S., W. Michalek, G. G. Presting, R. ten Hoopen, K. dos Santos *et al.*, 2001 Sequence organization of barley centromeres. Nucleic Acids Res. **29:** 5029–5035.

Hwang, C. K., E. S. Svarovskaia and V. K. Pathak, 2001 Dynamic copy choice: steady state between murine leukemia virus polymerase and polymerase-dependent RNase H activity determines frequency of in vivo template switching. Proc. Natl. Acad. Sci. USA **98:** 12209–12214.

Jaeger, J. A., D. H. Turner and M. Zuker, 1990 Predicting optimal and suboptimal secondary structure for RNA. Methods Enzymol. **183:** 281–306.

Jiang, N., Z. Bao, S. Temnykh, Z. Cheng, J. Jiang *et al.*, 2002a *Dasheng*: a recently amplified non-autonomous LTR element that is a major component of pericentromeric regions in rice. Genetics **161:** 1293–1305.

Jiang, N., K. Jordan and S. R. Wessler, 2002b *Dasheng* and *RIRE*2. A nonautonomous long terminal repeat element and its putative autonomous partner in the rice genome. Plant Physiol. **130:** 1697–1705.

Kalendar, R., T. Grob, M. Regina, A. Suoniemi and A. H. Schulman, 1999 IRAP and REMAP: two new retrotransposon-based DNA fingerprinting techniques. Theor. Appl. Genet. **98:** 704–711.

Kalendar, R., J. Tanskanen, S. Immonen, E. Nevo and A. H. Schulman, 2000 Genome evolution of wild barley (*Hordeum spontaneum*) by *BARE*-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc. Natl. Acad. Sci. USA **97:** 6603–6607.

Kankaanpää, J., L. Mannonen and A. H. Schulman, 1996 The genome sizes of *Hordeum* species show considerable variation. Genome **39:** 730–735.

Kerwood, D. J., M. J. Cavaluzzi and P. N. Borer, 2001 Structure of SL4 RNA from the HIV-1 packaging signal. Biochemistry **40:** 14518–14529.

Kullback, S., and R. A. Leibler, 1951 On information and sufficiency. Ann. Math. Stat. **22:** 79–86.

Kumar, A., and J. Bennetzen, 1999 Plant retrotransposons. Annu. Rev. Genet. **33:** 479–532.

Kumekawa, N., H. Ohtsubo, T. Horiuchi and E. Ohtsubo, 1999 Identification and characterization of novel retrotransposons of the gypsy type in rice. Mol. Gen. Genet. **260:** 593–602.

Lucas, H., G. Moore, G. Murphy and R. B. Flavell, 1992 Inverted repeats in the long-terminal repeats of the wheat retrotransposon Wis-2-1A. Mol. Biol. Evol. **9:** 716–728.

Manninen, I., and A. H. Schulman, 1993 *BARE*-1, a *copia*-like retroelement in barley (*Hordeum vulgare* L.). Plant Mol. Biol. **22:** 829–846.

Manninen, O., R. Kalendar, J. Robinson and A. H. Schulman, 2000 Application of *BARE*-1 retrotransposon markers to map a major resistance gene for net blotch in barley. Mol. Gen. Genet. **264:** 325–334.

Marquet, R., C. Isel, C. Ehresmann and B. Ehresmann, 1995 tRNAs as primer of reverse transcriptases. Biochemie **77:** 113–124.

McAllister, B. F., and J. H. Werren, 1997 Phylogenetic analysis of a retrotransposon with implications for strong evolutionary constraints on reverse transcriptase. Mol. Biol. Evol. **14:** 69–80.

Meyers, B. C., S. V. Tingey and M. Morgante, 2001 Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res. **11:** 1660–1676.

Monfort, A., C. M. Vicient, R. Raz, P. Puigdomenech and J. A. Martinez-Izquierdo, 1995 Molecular analysis of a putative transposable retroelement from the Zea genus with internal clusters of tandem repeats. DNA Res. **2:** 255–261.

Peleg, O., S. Brunak, E. N. Trifonov, E. Nevo and A. Bolshoy, 2002 RNA secondary structure and sequence conservation in C1 region of human immunodeficiency virus type 1 env gene. AIDS Res. Hum. Retroviruses **18:** 867–878.

Petersen, G., and O. Seberg, 1997 Phylogenetic analysis of the Triticeae (Poaceae) based on rpoA sequence data. Mol. Phylogenet. Evol. **7:** 217–230.

Roebuck, K. A., and M. Saifuddin, 1999 Regulation of HIV-1 transcription. Gene Exp. **8:** 67–84.

Sambrook, J., E. F. Fritsch and T. Maniatis, 1989 *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

SanMiguel, P., A. Tikhonov, Y. K. Jin, N. Motchoulskaia, D. Zakharov *et al.*, 1996 Nested retrotransposons in the intergenic regions of the maize genome. Science **274:** 765–768.

Schneider, T. D., 1999 Measuring molecular information. J. Theor. Biol. **201:** 87–92.

Schneider, T., and R. Stephens, 1990 Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. **18:** 6097–6100.

Schneider, T. D., G. D. Stormo, L. Gold and A. Ehrenfeucht, 1986 Information content of binding sites on nucleotide sequences. J. Mol. Biol. **188:** 415–431.

Schultz, S. J., M. Zhang, C. D. Kelleher and J. J. Champoux, 2000 Analysis of plus-strand primer selection, removal, and reutilization by retroviral reverse transcriptases. J. Biol. Chem. **275:** 32299–32309.

Shirasu, K., A. H. Schulman, T. Lahaye and P. Schulze-Lefert, 2000 A contiguous 66 kb barley DNA sequence provides evidence for reversible genome expansion. Genome Res. **10:** 908–915.

Siebert, P. D., A. Chenchik, D. E. Kellogg, K. A. Lukyanov and S. A. Lukyanov, 1995 An improved PCR method for walking in uncloned genomic DNA. Nucleic Acids Res. **23:** 1087–1088.

Suoniemi, A., K. Anamthawat-Jónsson, T. Arna and A. H. Schulman, 1996 Retrotransposon *BARE*-1 is a major, dispersed component of the barley (*Hordeum vulgare* L.) genome. Plant Mol. Biol. **30:** 1321–1329.

Suoniemi, A., D. Schmidt and A. H. Schulman, 1997 *BARE*-1 insertion site preferences and evolutionary conservation of RNA and cDNA processing sites. Genetica **100:** 219–230.

Suoniemi, A., J. Tanskanen and A. H. Schulman, 1998a Gypsy-like retrotransposons are widespread in the plant kingdom. Plant J. **13:** 699–705.

Suoniemi, A., J. Tanskanen, O. Pentikäinen, M. S. Johnson and A. H. Schulman, 1998b The core domain of retrotransposon integrase in *Hordeum*: predicted structure and evolution. Mol. Biol. Evol. **15:** 1135–1144.

Thompson, J. D., D. G. Higgins and T. J. Gibson, 1994 Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:** 4673–4680.

Vicient, C. M., A. Suoniemi, K. Anamthawat-Jónsson, J. Tanskanen, A. Beharav *et al.*, 1999 Retrotransposon *BARE*-1 and its role in genome evolution in the genus *Hordeum*. Plant Cell **11:** 1769–1784.

Vicient, C. M., M. Jääskeläinen, R. Kalendar and A. H. Schulman, 2001a Active retrotransposons are a common feature of grass genomes. Plant Physiol. **125:** 1283–1292.

Vicient, C. M., R. Kalendar and A. H. Schulman, 2001b Envelope-containing retrovirus-like elements are widespread, transcribed and spliced, and insertionally polymorphic in plants. Genome Res. **11:** 2041–2049.

Voytas, D. F., M. P. Cummings, A. K. Konieczny, F. M. Ausubel and S. R. Rodermel, 1992 *Copia*-like retrotransposons are ubiquitous among plants. Proc. Natl. Acad. Sci. USA **89:** 7124–7128.

Wicker, T., N. Stein, L. Albar, C. Feuillet, E. Schlagenhauf *et al.*, 2001 Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum*) reveals multiple mechanisms of genome evolution. Plant J. **26:** 307–316.

Witte, C. P., Q. H. Le, T. Bureau and A. Kumar, 2001 Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proc. Natl. Acad. Sci. USA **98:** 13778–13783.

Xiong, Y., and T. H. Eickbush, 1990 Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. **9:** 3353–3362.

Yang, J., H. Bogerd, S. Y. Le and B. R. Cullen, 2000 The human endogenous retrovirus K Rev response element coincides with a predicted RNA folding region. RNA **6:** 1551–1564.

Zeeberg, B, 2002 Shannon information theoretic computation of synonymous codon usage biases in coding regions of human and mouse genomes. Genome Res. **12:** 944–955.

Zuker, M., 1989 On finding all suboptimal foldings of an RNA molecule. Science **244:** 48–52.

Zuker, M., 1994 Prediction of RNA secondary structure by energy minimization. Methods Mol. Biol. **25:** 267–294.

Zuker, M., and P. Stiegler, 1981 Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information. Nucleic Acids Res. **9:** 133–148.

Communicating editor: D. F. Voytas