Highly abundant pea LTR retrotransposon *Ogre* is constitutively transcribed and partially spliced

Pavel Neumann*, Dana Požárková and Jiří Macas
Institute of Plant Molecular Biology, Laboratory of Molecular Cytogenetics, Branišovská 31, České Budějovice, 37005 Czech Republic (*author for correspondence; e-mail neumann@umbr.cas.cz)

Abstract

We have isolated and characterized a novel giant retroelement, named *Ogre*, which is over 22 kb long and makes up at least 5% of the pea (*Pisum sativum* L.) genome. This element can be classified as a Ty3/*gypsy*-like LTR retrotransposon based on the presence of long terminal repeats (LTRs) and the order of the domains coding for typical retrotransposon proteins. In addition to its extreme length, it has several features which make it unique among the retroelements described so far: (1) the sequences coding for *gag* and *prot* proteins are separated from the *rt/rh-int* domains by several stop codons; (2) the region containing these stop codons is removed from the element transcripts by splicing which results in reconstitution of the complete *gag-pol* coding sequence; (3) only a part of the transcripts is spliced which probably determines the ratio of translated proteins; (4) the element contains an extra ORF located upstream the *gag-pol* coding sequences, potentially coding for a protein of 546–562 amino acids with unknown function. The transcriptional activity of the *Ogre* elements has been detected in all organs tested (leaves, roots, flowers) as well as in wounded leaves and protoplasts. Considering this retroelement's constitutive expression and observed high mutual similarity of the element genomic sequences, it is possible to speculate about its recent amplification in the genomes of pea and other legume plants.

Introduction

Retrotransposons represent a class of mobile elements characterized by a replicative mode of transposition. Their replication cycle includes transcription into an RNA intermediate, which is translated into the element-specific proteins facilitating its reverse transcription and subsequent integration into the genome. Therefore, contrary to DNA transposons employing an excision-reintegration mechanism of transposition, the mobility of retrotransposons is always accompanied by an increase in their copy number in the genome. Retrotransposons are widespread among eukaryots (Hirochika and Hirochika, 1993; Bennetzen, 1996; Suoniemi *et al.*, 1998; Kumekawa *et al.*, 1999b;

Schmidt, 1999; Arkhipova, 2001; Bhattacharya *et al.*, 2002; Deininger and Batzer, 2002; Wöstemeyer and Kreibich, 2002) and, as a consequence of their replication, they often make up a considerable portion of eukaryotic genomes (SanMiguel *et al.*, 1996; Shirasu *et al.*, 2000; Deininger and Batzer, 2002). In plants, the copy number of elements belonging to individual retrotransposon families vary greatly, from a few up to 10^4–10^5 copies per haploid genome (1C) (Bennetzen, 1996; SanMiguel and Bennetzen, 1998; Kumar and Bennetzen, 1999; Feschotte *et al.*, 2002). The variability in retrotransposon accumulation is supposed to significantly contribute to observed differences in genome sizes between even closely related plant species (SanMiguel *et al.*, 1996, 1998; Bennetzen, 1998; Vicient *et al.*, 1999; Kalendar *et al.*, 2000).

The retrotransposons are classified as LTR or non-LTR elements based on the presence of long terminal

repeats (LTRs) flanking the retroelement (Kumar and Bennetzen, 1999; Feschotte *et al.*, 2002). In plants, most of the currently described elements belong to the LTR retrotransposons, which include a large number of diverse families (Bennetzen, 1998). Their LTR sequences are hundreds to several thousand bases in length, carry regions important for transcription (promoter, polyadenylation signal), and play a key role in a strand transfer occurring during the element's reverse transcription (Temin, 1981; Kumar and Bennetzen, 1999; Bennetzen, 2000; Wilhelm and Wilhelm, 2001). In their vicinity there are additional sequences important for element replication: the primer binding site (PBS, complementary to the 3′ end of certain tRNA) downstream of the left LTR, and the polypurine tract (PPT) upstream of the right LTR (Temin, 1981; Kumar and Bennetzen, 1999; Wilhelm and Wilhelm, 2001). The region between the two LTRs contains two genes, *gag* and *pol*, encoding all proteins necessary for the element's transposition (Boeke and Stoye *et al.*, 1997). In contrast to retroviruses, retrotransposons are not infectious because they lack a functional *env* gene. The *gag* gene codes for a capside protein responsible for the formation of virus-like particles. The *pol* gene codes for a protease (*prot*), a reverse transcriptase/RNaseH (*rt/rh*) and an integrase (*int*) which are required for the replication of the retrotransposon RNA and integration of a new copy into the genome. These coding sequences are usually placed in one or two ORFs separated by a stop codon and translated into a single polyprotein (*gag-pol*) which is cleaved into individual proteins by the protease. Based on the order of domains encoded within the *pol* gene the LTR retrotransposons are classified either as Ty3/*gypsy* (*Metaviridae*) or Ty1/*copia*-like (*Pseudoviridae*). Whereas the former family has the same order of domains as retroviruses (*prot-rt/rh-int*), the latter one has the last two domains reversed (*prot-int-rt/rh*) (Kumar and Bennetzen, 1999; Hull, 2001). The structural difference of these two families is also reflected in their sequence divergence (Xiong and Eickbush, 1990; Malik and Eickbush, 2001).

In our previous work (Neumann *et al.*, 2001), we isolated several partial clones of highly abundant repetitive elements from the pea (*Pisum sativum* L.) genome, which were shown to be conserved in sequence and have a dispersed genomic organization. As some of these clones contained sequences not reported until now, we presumed that they may belong to a novel family of retroelement-like repeats. In this study we aimed to isolate and characterize the full-

length elements corresponding to one of these clones, Psat-17. Our results show that this sequence is part of a giant LTR-retrotransposon, named *Ogre*, which displays several features unique among all retroelements and constitutes a significant portion of the pea genome.

Materials and methods

Plant material

Seeds of pea (*Pisum sativum* L. cv. Carrera) were obtained from the Plant Breeding Station at Boršov, Czech Republic. Total genomic DNA was extracted from leaves as described by Dellaporta *et al.* (1983). All DNA concentration measurements were done with PicoGreen dye (Molecular Probes) according to the manufacturer's recommendations.

Genomic libraries

The pea lambda DASHII library was purchased from Stratagene. The cosmid library was constructed by cloning pea genomic DNA partially digested with *Mbo*I into SuperCos-1 vector (Stratagene) according to the manufacturer's protocol.

Both lambda and cosmid libraries were screened with the Psat17 probe (Neumann *et al.*, 2001), labelled and detected with the AlkPhos Direct Hybridization and Detection Kit (Amersham). Selected positive clones were subcloned in order to obtain templates suitable for shotgun sequencing. The subcloning included random fragmentation of purified DNA by sonication, generation of blunt ends with mung bean nuclease (New England Biolabs), and isolation of a fraction corresponding to 1200–1700 bp by preparative agarose gel electrophoresis. The DNA was then treated with polynucleotide kinase (Roche), ligated into *Sma*I-digested plasmid pBluescript II SK+ and transformed into XL1-Blue MRF' cells. The subcloned fragments were sequenced from both ends using the dideoxy-mediated chain-termination method (Sanger *et al.*, 1977) and the clone sequences were assembled with Staden Package software (Staden *et al.*, 1996).

Sequence analysis

Basic sequence analysis was done with Staden Package software (Staden *et al.*, 1996) and program tools implemented at the Biology Work-

Bench server (http://workbench.sdsc.edu/). Multiple sequence comparisons were performed with CLUSTALW (Thompson *et al.*, 1994) and plots of sequence conservation along alignments were generated by the PLOTCON program (included in the EMBOSS package, Rice *et al.*, 2000). Homology searches were performed with BLAST and FASTA (Altschul *et al.*, 1997, and Pearson and Lipman, 1988, respectively). The search for conserved protein domains was done with RPS-Blast (Marchler-Bauer *et al.*, 2003; http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). Sequences of tRNAs used for identification of PBS were obtained from the *Arabidopsis thaliana* tRNA database (Lowe and Eddy, 1997; http://rna.wustl.edu/tRNAdb/). Splice site analysis was performed at the NetGene2 server (Hebsgaard *et al.*, 1996; http://www.cbs.dtu.dk/services/NetGene2/). Potential transmembrane domains were predicted using TMAP (Persson and Argos, 1994; included in the EMBOSS package, Rice *et al.*, 2000), Tmpred (Hofman and Stoffel, 1993; http://www.ch.embnet.org/software/TMPRED_form.html), HMMTOP (Tusnády and Simon, 2001; http://www.enzim.hu/hmmtop/html/submit.html) and TMHMM (Krogh *et al.*, 2001; http://www.cbs.dtu.dk/services/TMHMM/).

RNA isolation and analysis

The tissues used for RNA isolation (leaf, root, flower) were harvested from plants cultivated in pots under a 15 h light (22 °C)/9 h dark (18 °C) photoperiod. Alternatively, wounded leaves were prepared by partial abrading a leaf surface of these plants with emery paper followed by plant incubation for 2 days before leaf harvesting. Protoplasts were prepared from leaves cut into smaller pieces and incubated in an enzyme mixture (0.5 M sucrose, 1.375% cellulase R10, 0.2% macerozyme R10, 5 mM $CaCl_2$, 10 mM K-MES) for 15 h at room temperature. Total RNA was isolated with the Total RNA Isolation Kit (Ambion). All RNA samples were treated with RNase-free DNase (Gibco-BRL) to remove any contaminant DNA. Reverse transcription was carried out with the SuperScript II Reverse Transcription Kit (Gibco-BRL) by a random priming method according to the manufacturer's recommendations, with 1 μg of the template RNA. The RT-PCR reaction mix (25 μl) consisted of 1× PCR buffer, 0.1 mM dNTPs, 0.2 μM primers, 1.5 mM $MgCl_2$, 1 U of Platinum *Taq* polymerase (Invitrogen) and 2.5 ng of

reverse-transcribed RNA or equal amount of reverse transcriptase-untreated RNA as a negative control. The following primers were used in the RT-PCR experiments: 7 (5′-CAGGACTCGTTCTACGAAGC-3′), 8 (5′-AGCAGAGCAGAATGGTCAAG-3′), 9 (5′-TTTCATTGGAGGAAATCCAG-3′), 10 (5′-TTCTCGCACTCCTCTGAGTT-3′), 11 (5′-GGCTCATCAACTCAGAGGAG-3′) and 12 (5′-CAGCCTCTTGTCTCTGAACC-3′). The reaction profile included 40 cycles of 30 s at 94 °C, 50 s at 55 °C and 1–3 min at 72 °C preceded by initial denaturation (3 min at 94 °C) and followed by final extension step (10 min at 72 °C). Reaction products were resolved by agarose gel electrophoresis.

Results

Ogre is an extremely long Ty3/gypsy-like element

Screening the pea genomic libraries cloned in lambda and cosmid vectors with the Psat 17 probe (Neumann *et al.*, 2001) revealed large numbers of positive clones, five of which (three lambda and two cosmid clones) were picked at random and sequenced. Analysis of the resulting sequences showed that all clones contained overlapping parts of the same retroelement and that, due to its large size, the full-length sequence of the element was only present in one of the clones (Pscos16, Figure 1C). The full-length element found in this clone is 22 753 bp long and is flanked by long terminal repeats of 5040 and 5217 bp (Figure 1A). Both LTRs terminate in short inverted repeats, TG at the 5′ end and CA at the 3′ end, and their different size is caused by an insertion/duplication of 177 bp within the right LTR. The element is bordered by 5 bp direct repeats, which probably represent a target site duplication generated during the element integration. The primer binding site showing complete identity to the 18 bp at the 3′ end of $tRNA_{arg}$ (tRNA28 and tRNA104 from the *Arabidopsis thaliana* tRNA database; Lowe and Eddy, 1997) was identified 3 bp downstream of the left LTR; this region is fully conserved among the sequences present in all five clones. A putative polypurine tract was located 9 bp upstream of the right LTR (Figure 1A). Searching the database of conserved domains (RPS-Blast) with *Ogre* sequences as queries allowed identification of the regions potentially coding for *gag*, *prot*, *rt/rh* and *int* proteins (Table 1 and Figure 1A). Based on the sequence homology and the order of the coding domains, the *Ogre* element could be classified as a LTR retrotransposon of
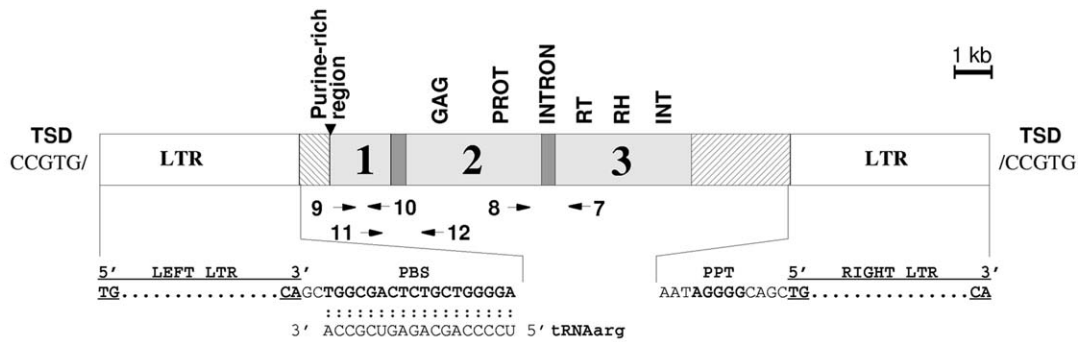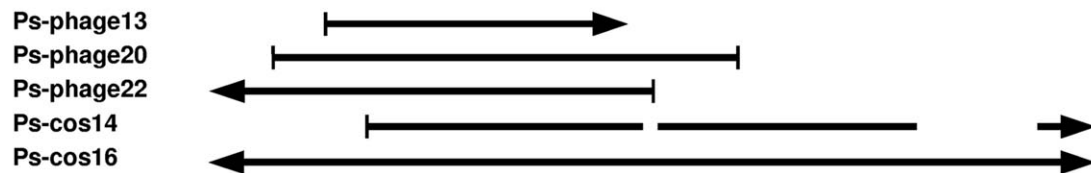
Figure 1. Schematic representation of *Ogre* element. A. Structure of the full-length element found in the clone Ps-cos16. The coding sequences ORF1, ORF2 and ORF3 are represented by light grey boxes marked with corresponding numbers. Dark grey boxes indicate regions separating individual ORFs, and cross-hatched boxes represent putative 5′- and 3′-untranslated regions (UTRs). The positions of individual coding domains within ORF2 and ORF3 are indicated above, together with the position of the intron between these ORFs and of the purine-rich region upstream of ORF1. Other structural or functional features are also shown and marked as follows: LTR, long terminal repeat; PBS, primer binding site (its complementarity to the 3′ end of tRNA$_{arg}$ is indicated); PPT, polypurine tract; TSD, target site duplication. The arrows numbered 7–12 mark positions and directions (5′→3′) of primers used for RT-PCR experiments. B. Positions of stop codons in the full-length *Ogre* sequence (clone Ps-cos16) plotted in three reading frames. Each vertical line represents a single stop codon present in the respective frame. C. A representation of *Ogre* sequences present in the five sequenced clones: Ps-phage13 (acc. no. AY299394), Ps-phage20 (AY299395), Ps-phage22 (AY299396), Ps-cos14 (AY299397), Ps-cos16 (AY299398). The lines show parts of the clones homologous to the element; dimension lines indicate ends of inserts, arrow lines indicate that inserts continue with a sequence not belonging to *Ogre*. The gaps in the Ps-cos14 line indicate missing parts of the sequence.

the Ty3/*gypsy* type (the family of *Metaviridae*; Hull, 2001).

The *Ogre* copies in the five sequenced clones showed high overall sequence identity, with the most conserved parts found in their potential coding regions (average identity 91% on the DNA and 92% on the protein level). The least similarity was detected among LTRs from different clones which varied in length (2873–5217 bp) as well as in their nucleotide sequence (average identity 80%). On the other hand, comparison of the two LTRs belonging to the same element present in the clone Ps-cos16 revealed that they are almost identical (99.6% similarity over 5217 bp). The similar result was obtained in case of the clone Ps-cos14, although one of its LTRs was truncated due

to cloning and thus only remaining 707 bp could be compared, showing 97.9% similarity.

Unique features of the Ogre *sequences: unusual arrangement of coding domains and the presence of an extra ORF*

Analysis of the *Ogre* sequences in all possible reading frames identified three non-overlapping ORFs, which we defined as regions devoid of stop codons (Figure 1B). These sequences are separated by short regions (283–374 bp between ORF1 and ORF2, and 243–244 bp between ORF2 and ORF3) with higher A/T content compared to the coding sequences (63% vs. 53% in ORFs) and containing stop codons that

Table 1. Conserved domains found in proteins encoded by *Ogre.*

ORF[a]	CD[b]	E-value	Position[c]	Active-site motif	Note
1	–	–	–		
2	*gag* (pfam03732)	$9e^{-10}$	240–331		the domain is not followed by zinc-finger motif CCHC; there are two alternative motifs, $CX_2HX_5HHX_3C$ and $CX_3CX_2CX_6C$, which differ from generalized zinc-finger motif $CX_{2-5}CX_{4-12}C/HX_{2-4}C/H$[h]
	prot (pfam00026)	5.2	813–953	DTG[d]	
3	*rt* (pfam00078)	$4e^{-18}$	1296–1465	YVDD[e]	
	rh (pfam00075)	$3e^{-06}$	1749–1845	DEDD[f]	
	int (pfam00665)	$7e-25$	2006–2162	DDE[g]	the domain is preceded by zinc-binding motif HHCC[g]

[a]Protein sequences deduced from ORFs of the *Ogre* element present in the clone Ps-cos16 were used to search conserved domain database (CDD) using RPS-Blast (Marchler-Bauer *et al.*, 2003). ORF2 and ORF3 were translated into one protein as it would be from the spliced transcript.

[b]Names and accession numbers of the conserved domains (CD) deposited in CDD.

[c]Positions of identified domains in the putative protein sequence.

[d]Skalka (1989).

[e]Xiong and Eickbush, 1990; Ding *et al.*, 1998.

[f]Davies *et al.*, 1991; Malik and Eickbush, 2001.

[g]Asante-Appiah and Skalka, 1997; Maignan *et al.*, 1998.

[h]Burd and Dreyfuss, 1994.

are highly conserved among all clones sequenced. Based on the homologies with the conserved protein sequences mentioned above, all domains typical of retroelements could be associated with ORF2 and ORF3 (Figure 1A, B). On the other hand, ORF1 did not show any homology to retroelements or any other sequences of known origin or function. The first starting codon within ORF1 is located 763–823 bp downstream of the left LTR and defines a potential coding sequence of 1638–1686 bp (i.e. 546–562 amino acids when translated to protein). This ORF is conserved in sequence and terminates at the same stop codon in all five sequenced clones and thus most likely represents an as of yet unknown type of retroelement gene. The ORF is preceded by a purine-rich region (92% A+G) of 62–70 bp located 688–725 bp downstream of the left LTR.

Although the order of domains coding for the retroelement proteins found in the *Ogre* sequences is typical of Ty3/*gypsy* retrotransposons, the presence of the stop codon-rich region separating the *gag-prot* and *rt/rh-int* coding sequences (Figure 1B) is a unique feature of this element. Positions of these stop codons are highly conserved between the clones (with the exception of Ps-cos14 which lacks the whole region), which rules out the possibility that they were generated by random mutations of an originally uninterrupted open reading frame. Such continuous open reading frames comprising all coding sequences are typical of most retroviruses and retrotransposons analysed so far, although they may contain a single stop codon between the *gag* and *pol* (*prot-rt/rh-int*) and/or between the *prot* and *rt/rh* domains (Vogt, 1997; Kumar and Bennetzen, 1999; Peterson-Burch and Voytas, 2002). This stop codon is then either misread or bypassed by a ribosomal frameshift in order to facilitate translation of the following domains (Gesteland and Atkins, 1996; Swanstrom and Wills, 1997; Alam *et al.*, 1999). However, such mechanisms cannot be expected to efficiently overcome several stop codons present in the *Ogre* sequences. These findings prompted us to perform a computer analysis of potential splice sites in order to see whether all stop codons between ORF2 and ORF3 could be removed from the *Ogre* transcripts in one turn via intron splicing. The analysis performed with the NetGene2 program (Hebsgaard *et al.*, 1996) provided strong support for this hypothesis, predicting donor and acceptor sites at high confidence ($P = 0.96$) surrounding the region containing the stop codons. On the other hand, the same test performed for the region separating ORF1 from ORF2 did not predict splicing of its sequence.

Ogre *sequences are constitutively transcribed and partially spliced*

The transcriptional activity of the *Ogre* elements was studied by RT-PCR with total RNA extracted from various tissues (leaves, roots, flowers) as a template.

In addition to the tissues harvested from plants grown in normal conditions, RNA prepared from wounded leaves and protoplasts was also tested, because these stress conditions are known to elicit transcription of many plant retroelements that are otherwise silent (Grandbastien, 1998). The primers derived from ORF1 (Figure 1A, primers 9 and 10) were used for the reaction, taking advantage of the specificity of this sequence for the *Ogre* elements. Surprisingly, the fragments of expected size were amplified not only from the samples subjected to stress, but also from the tissues harvested from untreated plants (Figure 2A). These experiments thus showed that *Ogre* elements belong to the very few plant retrotransposons that are constitutively expressed in the whole plants.

The observed transcriptional activity also provided a possibility to test eventual splicing of the element's RNA. While the primers designed to amplify the region between ORF1 and ORF2 (primers 11 and 12) generated a single amplification product of the size corresponding to the full-length sequence (Figure 2B), the primers directed to the region between ORF2 and ORF3 (primers 7 and 8) produced two major products, one of the size of the full-length genomic DNA sequence (850 bp, see also line marked 'DNA' on Figure 2), and one shorter fragment of about 600 bp (Figure 2C). These results are in agreement with the computer prediction of the intron only for the area between ORF2 and ORF3 but not for the region between ORF1 and ORF2. To further verify these results we cloned and sequenced both size variants (full-length and potentially spliced) generated in RT-PCR with primers 7 and 8. The sequence comparisons revealed that the shorter fragments lacked the region exactly between the predicted donor and acceptor sites, thus confirming their origin by splicing (Figure 3). Interestingly, only a part of the *Ogre* transcripts seemed to be spliced as demonstrated by the presence of the two bands. This analysis also showed that the sequence in the clone Ps-cos14 probably represents a spliced copy of the element reintegrated into the genome, because it lacks the intron sequence. In order to estimate the proportion of full-length to spliced elements in the pea genome, we have sequentially screened phage and cosmid libraries with probes for the intron region including adjacent ORF2 sequences or for the intron sequence alone. The frequencies of clones simultaneously hybridizing to these two probes and to the former probe alone were similar in both libraries and showed that the ratio of unspliced to spliced *Ogre* sequences in the genome is about 65:1.
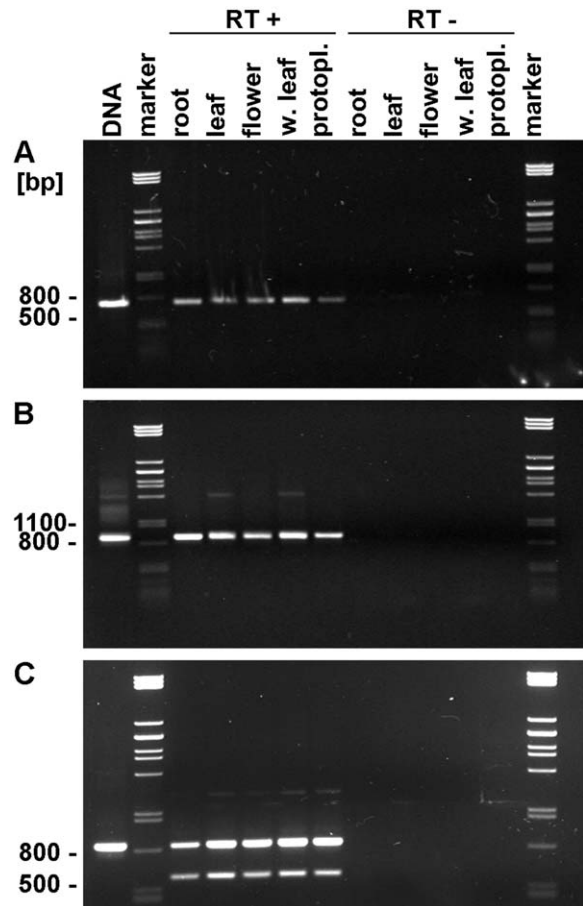

Figure 2. Transcriptional activity and splicing of *Ogre* sequences analysed by RT-PCR. The reactions were performed with total RNA isolated either from tissues harvested from plants grown at normal conditions (root, leaf, flower), or from tissues or cells subjected to stress conditions (wounded leaves and isolated protoplasts). Two sets of otherwise identical reactions differing only in including (RT+) or omitting (RT–) the reverse transcription step were always run in order to check for eventual false-positive results caused by contamination with genomic DNA. The primers were directed to amplify the following portions of *Ogre* transcripts (also indicated on Figure 1A). A. The ORF1 (primers 9 and 10). B. The region between ORF1 and ORF2 (primers 11 and 12). C. The region including putative intron between ORF2 and ORF3 (primers 7 and 8). The lines marked 'DNA' contain products of PCR reactions with pea genomic DNA and corresponding primers as a positive controls, 'marker' is lambda DNA digested with *Pst*I.

The elements are highly abundant in the pea genome and their homologues are also present in other legume species

In order to estimate the copy number of *Ogre* elements in the pea genome, known numbers of phage and cosmid pea genomic clones were screened with probes derived from defined regions of the element
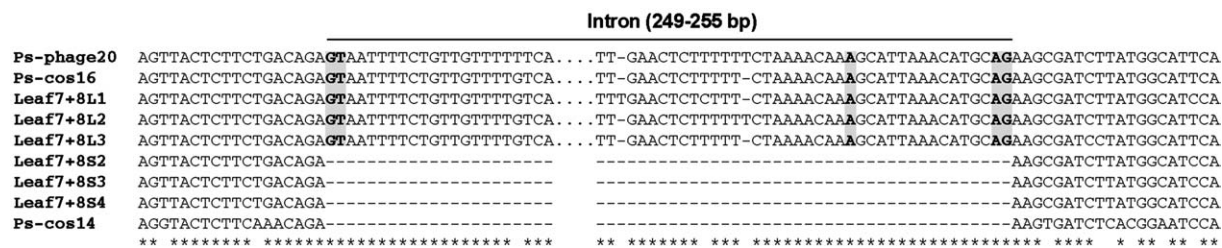
Intron (249-255 bp)

```
Ps-phage20   AGTTACTCTTCTGACAGAGTAATTTTCTGTTGTTTTTTCA....TT-GAACTCTTTTTTCTAAAACAAAGCATTAAACATGCAGAAGCGATCTTATGGCATTCA
Ps-cos16     AGTTACTCTTCTGACAGAGTAATTTTCTGTTGTTTTGTCA....TT-GAACTCTTTTT-CTAAAACAAAGCATTAAACATGCAGAAGCGATCTTATGGCATTCA
Leaf7+8L1    AGTTACTCTTCTGACAGAGTAATTTTCTGTTGTTTTGTCA....TTTGAACTCTCTTT-CTAAAACAAAGCATTAAACATGCAGAAGCGATCTTATGGCATCCA
Leaf7+8L2    AGTTACTCTTCTGACAGAGTAATTTTCTGTTGTTTTGTCA....TT-GAACTCTTTTTTCTAAAACAAAGCATTAAACATGCAGAAGCGATCTTATGGCATTCA
Leaf7+8L3    AGTTACTCTTCTGACAGAGTAATTTTCTGTTGTTTTGTCA....TT-GAACTCTTTTT-CTAAAACAAAGCATTAAACATGCAGAAGCGATCTTATGGCATTCA
Leaf7+8S2    AGTTACTCTTCTGACAGA--------------------------------------------------------------AAGCGATCTTATGGCATCCA
Leaf7+8S3    AGTTACTCTTCTGACAGA--------------------------------------------------------------AAGCGATCTTATGGCATCCA
Leaf7+8S4    AGTTACTCTTCTGACAGA--------------------------------------------------------------AAGCGATCTTATGGCATCCA
Ps-cos14     AGGTACTCTTCAAACAGA--------------------------------------------------------------AAGTGATCTCACGGAATCCA
             ** ******* *********************** ***    ** ******* *** ************************* ****  *  ** ** **
```

Figure 3. Sequence analysis of spliced transcripts. Multiple alignment of sequences surrounding the intron boundaries in three genomic clones (Ps-phage20: 10845–11135; Ps-cos14: 6882–6919; Ps-cos16: 25770–26059) and six RT-PCR products amplified from leaf RNA with primers 7 and 8 (Figure 2C). The RT-PCR products represent both unspliced (Leaf7+8L1 to 3) and spliced (Leaf7+8S2 to 4) variants of the transcripts. An average identity among sequences of genomic clones and RT-PCR products was 94% along the whole length of the amplified region and 96% in the intron. Intron terminal dinucleotides and putative branch point A are shaded; the gaps in aligned sequences are marked by dashes. Dots indicate inner parts of intron sequences which are not shown.

sequence. Four different probes, representing LTR, ORF1, ORF2, and ORF3 regions, respectively, were separately used for hybridizations. The proportion of positive clones depended on the probe, being about three times higher for LTR and ORF3 probes (11% of phage and 27% of cosmid clones) than for ORF1 and ORF2 probes (4% of phage and 8% of cosmid clones). Taking into account the pea genome size (1C = 4337 Mb, Baranyi and Greilhuber, 1996), the average insert size in the libraries (15 and 35 kb in the phage and cosmid library, respectively) and considering that each positive clone contained only a single copy of the element, the estimated numbers of copies per haploid genome were ca. 3×10^4 for LTR and ORF3, and 1×10^4 for ORF1 and ORF2, respectively.

Homology searches between *Ogre* sequences and other DNA and protein sequences were performed with a variety of programs employing Fasta and Blast algorithms. As expected, the regions ORF2 and ORF3 coding for typical retroelement protein domains showed high identities with a number of retroelement sequences from various plant species (data not shown). However, sequences homologous to regions which are not conserved due to their function, and thus could be considered *Ogre*-specific, were also found, predominantly in other legume plants (*Fabaceae*). The most interesting of these were the sequences homologous to the ORF1 region, which included partial clones of repetitive sequences from *Medicago sativa* (accession number L39961, Xia *et al.*, 1996), *Vicia faba* (accession number M16856, Kato *et al.*, 1985; AJ011710, Nouzová *et al.*, 1999) and *V. narbonensis* (accession number AJ391771, Nouzová *et al.*, 2001). The identification of these sequences suggests that *Ogre*-like elements might also occur in other legumes outside the *Pisum* genus. This hypothesis is also supported by

the detection of similar elements in a recently released *Medicago truncatula* working draft sequences (data not shown). These elements appear to have a sequence organization very similar to *Ogre*, and show sequence homology to all *Ogre* coding regions including the ORF1 (Figure 4B).

Discussion

The *Ogre* sequences reported in this paper have several features which make them unique among known retroelements. The complete *Ogre* element found in the clone Ps-cos16 is more than 22 kb long which by far exceeds the size range of plant retrotransposons reported to date (3–15 kb; Kumar and Bennetzen, 1999). This is caused in part by atypically long LTRs, spanning over 10 kb or 45% of the element sequence, whereas most other retrotransposons possess LTRs of hundreds rather than thousands of base pairs (Bennetzen, 1996; Kumar and Bennetzen, 1999). Moreover, the *Ogre* elements contain an extra ORF, which together with the relatively long regions between LTRs and coding sequences (5′- and 3′-untranslated regions) account for their exceptional size.

The presence of the extra ORF (ORF1, Figure 1A) of unknown origin and function is another remarkable feature of the *Ogre* sequences. This sequence could have been acquired by the element via capture of a cellular gene, as was reported for retroviruses (Telesnitsky and Goff *et al.*, 1997) and also for the maize retrotransposon *Bs1* (Elrouby and Bureau, 2001). However, the copy numbers of the ORF1 sequences in the pea genome are close to those of other parts of the *Ogre* elements, suggesting that they are present in most or all element copies. Additionally, a similar ORF1 region is present in *Ogre*-like ele-
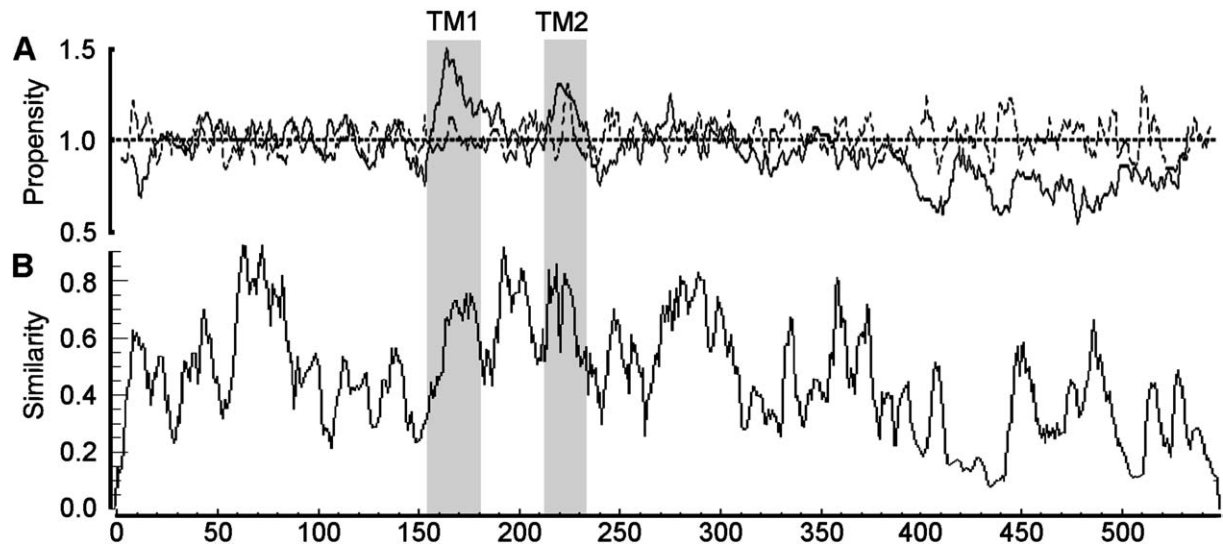
Figure 4. Prediction of transmembrane domains in the protein putatively encoded by ORF1. The prediction was done for a multiple alignment of ten protein sequences obtained by translation of ORF1 from the five *Ogre* clones described in this work and from five homologous *Ogre*-like elements identified in genomic sequences of *Medicago truncatula* (accession numbers AC135162.14, AC136506.8, AC126790.17, AC135315.6 and AC121234.13). A. A plot of potential transmembrane segments identified with the TMAP program (Persson and Argos, 1994). Propensity values above 1.0 indicate that the corresponding part of the protein is preferentially located in a lipid-associated structure. Two putative transmembrane segments, TM1 and TM2, are indicated by grey boxes, although the TM2 segment was predicted only for *Ogre* sequences from *P. sativum*. It should also be noted that positions of transmembrane segments in individual sequences predicted by other programs (HMMTOP and TMpred) slightly differed from those predicted by TMAP, but all of them were within more conserved regions between the residues 150 and 240. B. Graphical representation of the similarity along the multiple alignment used for the prediction of transmembrane domains displayed in panel A. The average similarities among the aligned sequences at each position were calculated for a window of 8 residues with the PLOTCON program.

ments which we have identified in the working draft sequences of *Medicago truncatula*. Therefore, it is likely that ORF1 is an integral part of the element and that it evolved or was acquired earlier in the evolution of the *Fabaceae* family.

The protein potentially encoded by the ORF1 displayed no similarity to the sequences characterized so far, so we can only speculate about its eventual function. Other extra ORFs of similar size have been identified downstream of the *gag/pol* gene in the plant retroelements *Cyclops-2*, *Calypso*, *Athila-4*, *Bagy-2*, *Rigy-2*, *Grande1*, *SIRE-1*, *Endovir* and *ToRTL1*, and it was proposed that, with the exception of *Grande1*, they code for *env*-like proteins (Martínez-Izquierdo *et al.*, 1997; Chavanne *et al.*, 1998; Late *et al.*, 1998; Laten, 1999; Peterson-Burch *et al.*, 2000; Vicient *et al.*, 2001b; Wright and Voytas, 2001). However, these sequences do not have homology to the ORF1 of *Ogre* (data not shown). On the other hand, the *env* proteins may be quite divergent in sequence (Lerat and Capy, 1999) and are therefore better identified by the presence of transmembrane domain which is crucial for their function (Swanstrom and Wills, 1997;

Vogt *et al.*, 1997). Interestingly, two transmembrane segments were predicted with the TMAP program (Persson and Argos, 1994) in the region of the *Ogre* ORF1 sequence which is conserved among the sequenced clones (Figure 4A). Their lengths were 29 and 20 residues, respectively, corresponding to the length of transmembrane domains of retroviruses as well as plant endogenous retroviruses (Bénit *et al.*, 2001; Wright and Voytas, 2001). Moreover, one of the transmembrane segments was also identified in corresponding regions of *Ogre*-like elements from *Medicago truncatula*. Although these sequences share only 34% identity to *Ogre* ORF1 proteins, the region around the transmembrane domains shows higher similarity, suggesting a conserved function (Figure 4). It should be noted, however, that an alternative program, TMHMM (Krogh *et al.*, 2001), did not predict these domains in pea or *M. truncatula* elements. ORFs coding for the *env* proteins of retroviruses and plant retrovirus-like elements are always located downstream the *gag-pol* sequences (Boeke and Stoye, 1997; Vogt *et al.*, 1997; Peterson-Burch *et al.*, 2000) which is just opposite to the ORF1 location in the *Ogre* elements. Extra

ORFs of unknown function located upstream of the *gag-pol* region have only been reported for rice retroelements *RIRE3*, *RIRE8A* and *RIRE8B* (Kumekawa *et al.*, 1999a). However, these ORFs are about half the size of ORF1, bear no detectable homology to *Ogre* elements and do not contain potential transmembrane domains (data not shown). An ORF upstream of the *gag* region and without any determined function was also reported for retrovirus WDSV, making it unique among all known retroviruses (Holzshu *et al.*, 1995).

Although *Ogre* can be classified as Ty3/*gypsy* element based on the order of its protein coding domains, the separation of the *gag-prot* region (ORF2) from the *rt/rh-int* region (ORF3, Figure 1B) by several stop codons and the removal of these stop codons by splicing have not been reported for any other retroelement. Splicing of the plant LTR retrotransposon was reported also for barley element *Bagy-2* in which it appears to ensure expression of the *env* gene. However, due to the presence of the *env* gene, the *Bagy-2* element is believed to be a plant endogenous retrovirus rather than a retrotransposon (Vicient *et al.*, 2001b). We presume that the splicing of the *Ogre* sequence may have two functions: (1) it allows efficient translation of the ORF3 region which can be read in frame with the ORF2 sequence in the spliced transcripts, and (2) it affects the ratio of proteins encoded by ORF2 and ORF3. As the observed splicing was only partial in all tissues tested (Figure 2C), it can be assumed that the *gag-prot* product is generally overproduced compared to *rt/rh-int*. Thus, splicing appears to be an alternative mechanism to the stop codon misreading or ribosomal frameshift which provides regulation of *gag* to *gag-pol* (*prot-rt/rh-int*) transcript ratios in some retroelements (Swanstrom and Wills, 1997; Vogt *et al.*, 1997; Shehu-Xhilaga *et al.*, 2001, and references therein).

An important question is what mechanism of translation initiation is used by *Ogre*, because the region upstream of ORF1 contains many start codons in all three reading frames and thus ribosome scanning of RNA from the cap to the first start codon in a favourable context seems to be unlikely. Several mechanisms including leaky scanning, translation re-initiation and cap-independent translation from the internal ribosome entry site (IRES) can be considered (Swanstrom and Wills, 1997). As the purine-rich region preceding the ORF1 strongly resembles the IRES recently described by Dorokhov *et al.* (2002), the latter mechanism appears to be most likely. This hypothesis is also supported by the presence of a similar region located just upstream of the first ORF in the retrotransposon *Cereba* (accession number AY040833) (data not shown).

The observed constitutive transcriptional activity is yet another surprising feature of *Ogre* elements. Although some recent papers describe transcriptional activity of retrotransposons in somatic cells (Rossi *et al.*, 2001; Vicient *et al.*, 2001a; Echenique *et al.*, 2002), it contrasts with many plant retrotransposons, which are transcriptionally silent or are transcriptionally active only during certain stages of plant development (Pouteau *et al.*, 1991; Suoniemi *et al.*, 1996; Turcich *et al.*, 1996) or stress conditions (Hirochika, 1993; Pearce *et al.*, 1996; Grandbastien *et al.*, 1997; Takeda *et al.*, 1998; Kato *et al.*, 1999; Beguiristain *et al.*, 2001). As the transposition of retrotransposons is believed to be regulated mainly on the transcriptional level, such limited activity might evolve in order to suppress deleterious effects of element amplification in the genome (Kumar and Bennetzen, 1999). The absence of this regulation favours the hypothesis of recent amplification of *Ogre* in pea where it constitutes over 5% of the genome. This hypothesis is also supported by high mutual similarities of *Ogre* sequences as well as the finding of highly identical LTRs within the elements. As the two LTRs are created by a duplication of a single template during the element's replication and then accumulate random mutations, the degree of their divergence is proportional to the time passed since the element's transposition (SanMiguel *et al.*, 1998). The small proportion of mutations (18 single-base substitutions and one 177 bp insertion/duplication within 5217 bp of LTRs in Pscos16) thus suggests relatively recent transposition of the respective element. On the other hand, we cannot rule out a possibility that *Ogre*'s transposition is blocked or at least regulated at some later point of its replication cycle.

Taken together, our data indicate that *Ogre* represents a novel family of retrotransposons, which is defined by the unique arrangement and splicing of the *gag-pol* coding domains, the presence of the extra ORF, the unusual PBS identical to the 3′ end of tRNA$_{arg}$ and the extreme size of the elements. Based on these features, the element cannot be simply assigned to either of the two existing genera within the *Metaviridae* family. Therefore, future research should be directed to functional, evolutionary and comparative analysis of the *Ogre* sequences to help us in deciding whether the elements constitute a novel taxonomic group. Moreover, the high overall sequence similarity,

conservation of coding sequences and transcriptional activity of known *Ogre* elements may indicate their recent transpositional activity, which makes them of great interest for studies of retrotransposon interaction with the host genomes and of the impact of their amplification on the genome size.

Acknowledgements

References

Alam, S.L., Atkins, J.F. and Gesteland, R.F. 1999. Programmed ribosomal frameshifting: Much ado about knotting! Proc. Natl. Acad. Sci. USA 96: 14177–14179.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25: 3389–3402.

Arkhipova, I.R. 2001. Transposable element in the animal kingdom. Mol. Biol. 35: 157–167.

Asante-Appiah, E. and Skalka, A.M. 1997. Molecular mechanisms in retrovirus DNA integration. Antiviral Res. 36: 139–156.

Baranyi, M. and Greilhuber, J. 1996. Flow cytometric and Feulgen analysis of genome size variation in *Pisum*. Theor. Appl. Genet. 92: 297–307.

Beguiristain, T., Grandbastien, M.A., Puigdomenech, P. and Casacuberta, J.M. 2001. Three *Tnt1* subfamilies show different stress associated patterns of expression in tobacco. Consequences for retrotransposon control and evaluation in plants. Plant Physiol. 127: 212–221.

Bénit, L., Dessen, P. and Heidmann, T. 2001. Identification, phylogeny, and evolution of retroviral elements based on their envelope genes. J. Virol. 75: 11709–11719.

Bennetzen, J.L. 1996. The contributions of retroelements to plant genome organization, function and evolution. Trends Microbiol. 4: 347–353.

Bennetzen, J.L. 1998. The structure and evolution of angiosperm nuclear genomes. Curr. Opin. Plant Biol. 1: 103–108.

Bennetzen, J.L. 2000. Transposable element contributions to plant gene and genome evolution. Plant Mol. Biol. 42: 251–269.

Bhattacharya, S., Bakre, A. and Bhattacharya, A. 2002. Mobile genetic elements in protozoan parasites. J. Genet. 81: 73–86.

Boeke, J.D. and Stoye, J.P. 1997. Retrotransposons, endogenous retroviruses, and the evolution of retroelements. In: J.M. Coffin, S.H. Hughes and H.E. Varmus (Eds.) Retroviruses, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 343–436.

Burd, C.B. and Dreyfuss, G. 1994. Conserved structures and diversity of functions of RNA-binding proteins. Science 265: 615–621.

Chavanne, F., Zhang, D.X., Liaud, M.F. and Cerff, R. 1998. Structure and evolution of *Cyclops*: a novel giant retrotransposon of the Ty3/*gypsy* family highly amplified in pea and other legume species. Plant Mol. Biol. 37: 363–375.

Davies, J.F., Hostomska, Z., Hostomsky, Z., Jordan, S.R. and Matthews, D.A. 1991. Crystal structure of the ribonuclease H domain of HIV-1 reverse transcriptase. Science 252: 88–95.

Deininger, P.L. and Batzer, M.A. 2002. Mammalian retroelements. Genome Res. 12: 1455–1465.

Dellaporta, S.L., Wood, J. and Hicks, J.B. 1983. A plant DNA minipreparation: version II. Plant Mol. Biol. Rep. 1: 19–21.

Ding, J., Das, K., Hsiou, Y., Sarafianos, S.G., Clark, A.D., Jacobo-Molina, A., Tantillo, C., Hughes, S.H. and Arnold, E. 1998. Structure and functional implications of the polymerase active site region in a complex of HIV-1 RT with a double-stranded DNA template-primer and an antibody Fab fragment at 2.8 A resolution. J. Mol. Biol. 284: 1095–1111.

Dorokhov, Y.L., Skulachev, M.V., Ivanov, P.A., Zvereva, S.D., Tjulkina, L.G., Merits, A., Gleba, Y.Y., Hohn, T. and Atabekov, J.G. 2002. Polypurine (A)-rich sequences promote cross-kingdom conservation of internal ribosome entry. Proc. Natl. Acad. Sci. USA 99: 5301–5306.

Echenique, V., Stamova, B., Wolters, P., Lazo, G., Carollo, V.L. and Dubcovsky, J. 2002. Frequencies of Ty1-*copia* and Ty3-*gypsy* retroelements within the *Triticae* EST databases. Theor. Appl. Genet. 104: 840–844.

Elrouby, N. and Bureau, T.E. 2001. A novel hybrid open reading frame formed by multiple cellular gene transductions by a plant long terminal repeat retroelement. J. Biol. Chem. 276: 41963–41968.

Feschotte, C., Jiang, N. and Wessler, S.R. 2002. Plant transposable elements: where genetics meets genomics. Genetics 3: 329–341.

Gesteland, R.F. and Atkins, J.F. 1996. Recoding: dynamic reprogramming of translation. Annu. Rev. Biochem. 65: 741–768.

Grandbastien, M.A. 1998. Activation of plant retrotransposons under stress conditions. Trends Plant Sci. 3: 181–187.

Grandbastien, M.A., Lucas, H., Morel, J.B., Mhiri, C., Vernhettes, S. and Casacuberta, J.M. 1997. The expression of the tobacco *Tnt1* retrotransposon is linked to plant defense responses. Genetica 100: 241–252.

Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J., Rouzé, P. and Brunak, S. 1996. Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. Nucl. Acids Res. 24: 3439–3452.

Hirochika, H. 1993. Activation of tobacco retrotransposons during tissue culture. EMBO J. 12: 2521–2528.

Hirochika, H. and Hirochika, R. 1993. Ty1-*copia* group retrotransposons as ubiquitous components of plant genomes. Jpn. J. Genet. 68: 35–46.

Hofmann, K. and Stoffel, W. 1993. TMBASE – a database of membrane spanning protein segments. Biol. Chem. Hoppe-Seyler 374: 166.

Holzschu, D.L., Martineau, D., Fodor, S.K., Vogt, V.M., Bowser, P.R. and Casey, J.W. 1995. Nucleotide sequence and protein analysis of a complex piscine retrovirus, Walley dermal sarcoma virus. J. Virol. 69: 5320–5331.

Hull, R. 2001. Classifying reverse transcribing elements: a proposal and a challenge to the ICTV. Arch. Virol. 146: 2255–2261.

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E. and Schulman, A.H. 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.

Kato, A., Iida, Y., Yakura, K. and Tanifuji, S. 1985. Sequence analysis of *Vicia faba* highly repeated DNA: the *Bam*HI repeated sequence families. Plant Mol. Biol. 5: 41–53.

Kato, H., Sriprasertsak, P., Seki, H., Ichinose, Y., Shiraishi, T. and Yamada, T. 1999. Functional analysis of retrotransposons in pea. Plant Cell Physiol. 40: 933–941.

Krogh, A., Larrson, B., von Heijne, G. and Sonnhammer, E.L.L. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305: 567–580.

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

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

Kumekawa, N., Ohtsubo, E. and Ohtsubo, H. 1999b. Identification and phylogenetic analysis of *gypsy*-type retrotransposons in the plant kingdom. Genes Genet. Syst. 74: 299–307.

Laten, H.M. 1999. Phylogenetic evidence for Ty1-*copia*-like endogenous retroviruses in plant genomes. Genetica 107: 87–93.

Laten, H.M, Majumdar, A. and Gaucher, E.A. 1998. *SIRE-1*, a *copia*/Ty1-like retroelement from soybean, encodes a retroviral envelope like protein. Proc. Natl. Acad. Sci. USA 95: 6897–6902.

Lerat, E. and Capy, E. 1999. Retrotransposons and retroviruses: analysis of the envelope gene. Mol. Biol. Evol. 16: 1198–1207.

Lowe, T.M. and Eddy, S.R. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucl. Acids Res. 25: 955–964.

Maignan, S., Guilloteau, J.P., Zhou-Liu, Q., Clément-Mella, C. and Mikol, V. 1998. Crystal structures of the catalytic domain of HIV-1 integrase free and complexed with its metal cofactor: high level of similarity of the active site with other viral integrases. J. Mol. Biol. 282: 359–368.

Malik, H.S. and Eickbush, T.H. 2001. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. Genome Res. 11: 1187–1197.

Marchler-Bauer, A. anderson, J.B., DeWeese-Scott, C., Fedorova, N.D., Geer, L.Y., He, S., Hurwitz, D.I., Jackson, J.D., Jacobs, A.R., Lanczycki, C.J., Liebert, C.A., Liu, C., Madej, T., Marchler, G.H., Mazumder, R., Nikolskaya, A.N., Panchenko, A.R., Rao, B.S., Shoemaker, B.A., Simonyan, V., Song, J.S., Thiessen, P.A., Vasudevan, S., Wang, Y., Yamashita, R.A., Yin, J.J. and Bryant, S.H. 2003. CDD: a curated Entrez database of conserved domain alignments. Nucl. Acids Res. 31: 383–387.

Martínez-Izquierdo, J.A., Garcia-Martínez, J. and Vicient, C.M. 1997. What makes *Grande1* retrotransposon different? Genetica 100: 15–28.

Neumann, P., Nouzová, M. and Macas, J. 2001. Molecular and cytogenetic analysis of repetitive DNA in pea (*Pisum sativum* L.). Genome 44: 716–728.

Nouzová, M., Kubaláková, M., Doleželová, M., Koblížková, A., Neumann, P., Doležel, J. and Macas, J. 1999. Cloning and characterization of new repetitive sequences in field bean (*Vicia faba* L.). Ann. Bot. (London) 83: 535–541.

Nouzová, M., Neumann, P., Navrátilová, A., Galbraith, D.W. and Macas, J. 2001. Microarray-based survey of repetitive genomic sequences in *Vicia* spp. Plant Mol. Biol. 45: 229–244.

Pearce, S.R., Kumar, A. and Flavell, A.J. 1996. Activation of the Ty1-*copia* group retrotransposons of potato (*Solanum tuberosum*) during protoplast isolation. Plant Cell Rep. 15: 949–953.

Pearson, W.R. and Lipman, D.J. 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85: 2444–2448.

Persson, B. and Argos, P. 1994. Prediction of transmembrane segments in proteins utilizing multiple sequence alignments. J. Mol. Biol. 237: 182–192.

Peterson-Burch, B.D. and Voytas, D.F. 2002. Genes of the *Pseudoviridae* (Ty1/*copia* retrotransposons). Mol. Biol. Evol. 19: 1832–1845.

Peterson-Burch, B.D., Wright, D.A., Laten, H.M. and Voytas, D.F. 2000. Retroviruses in plants? Trends Genet. 16: 151–152.

Pouteau, S., Huttner, E., Grandbastien, M.A. and Caboche, M. 1991. Specific expression of the tobacco *Tnt1* retrotransposon in protoplasts. EMBO J. 10: 1911-1918.

Rice, P., Longden, I. and Bleasby, A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 16: 276–277.

Rossi, M., Araujo, P.G. and Van Sluys, M. 2001. Survey of transposable elements in sugarcane expressed sequence tags (ESTs). Genet. Mol. Biol. 24: 147–154.

Sanger, F., Nicklen, D. and Coulson, A.R. 1977. DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. USA 74: 5463–5467.

SanMiguel, P. and Bennetzen, J. 1998. Evidence that recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Ann. Bot. 82: 37–44.

SanMiguel, P., Tikhonov, A., Jin, Y., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z. and Bennetzen, J.L. 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science 274: 765–768.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. and Bennetzen, J.L. 1998. The paleontology of intergene retrotransposons of maize. Nature Genet. 20: 43–45.

Schmidt, T. 1999. LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes. Plant. Mol. Biol. 40: 903–910.

Shehu-Xhilaga, M., Crowe, S.M. and Mak, J. 2001. Maintenance of the gag/gag-pol ratio is important for human immunodeficiency virus type 1 RNA dimerization and viral infectivity. J. Virol. 75: 1834–1841.

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

Skalka, A.M. 1989. Retroviral proteases: First glimpses at the anatomy of processing machine. Cell 56: 911–913.

Staden, R. 1996. The Staden sequence analysis package. Mol. Biotechnol. 5: 233–241.

Suoniemi, A., Narvanto, A. and Schulman, A.H. 1996. The *BARE-1* retrotransposon is transcribed in barley from an LTR promoter active in transient assays. Plant Mol. Biol. 31: 295–306.

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

Swanstrom, R. and Wills, J.W. 1997. Synthesis, assembly, and processing of viral proteins. In: J.M. Coffin, S.H. Hughes and H.E. Varmus (Eds.) Retroviruses, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 263–334.

Takeda, S., Sugimoto, K., Otsuki, H. and Hirochika, H. 1998. Transcriptional activation of the tobacco retrotransposon Tto1 by wounding and methyl jasmonate. Plant Mol. Biol. 36: 365–376.

Telesnitsky, A. and Goff, S.P. 1997. Reverse transcriptase and the generation of retroviral DNA. In: J.M. Coffin, S.H. Hughes and

H.E. Varmus (Eds.) Retroviruses, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 121–160.

Temin, H.M. 1981. Structure, variation and synthesis of retrovirus long terminal repeat. Cell 27: 1–3.

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

Turcich, M.P., Bokhari-Riza, A., Hamilton, D.A., He, C., Messier, W., Stewart, C.B. and Mascarenhas, J.P. 1996. *PREM-2*, a *copia*-type retroelement in maize is expressed preferentially in early microspores. Sex. Plant Reprod. 9: 65–74.

Tusnády, G.E. and Simon, I. 2001. The HMMTOP transmembrane topology prediction server. Bioinformatics 17: 849–850.

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

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

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

Vogt, V.M. 1997. Retroviral virions and genomes. In: J.M. Coffin, S.H. Hughes and H.E. Varmus (Eds.) Retroviruses, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 27–70.

Wilhelm, M. and Wilhelm, F.-X. 2001. Reverse transcription of retroviruses and LTR retrotransposons. Cell. Mol. Life Sci. 58: 1246–1262.

Wöstemeyer, J. and Kreibich, A. 2002. Repetitive DNA elements in fungi (Mycota): impact on genomic architecture and evolution. Curr. Genet. 41: 189–198.

Wright, D.A. and Voytas, D.F. 2001. *Athila4* of *Arabidopsis* and *Calypso* of soybean define a lineage of endogenous plant retroviruses. Genome Res. 12: 122–131.

Xia, X., Du, S. and Erickson, L. 1996. A moderately repetitive DNA sequence in alfalfa is transcribed in a floral-specific manner. Genome 39: 9–16.

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