Heredity (2009) 103, 157–167

www.nature.com/hdy

ORIGINAL ARTICLE

Evolutionary conserved lineage of *Angela*-family retrotransposons as a genome-wide microsatellite repeat dispersal agent

P Smýkal[1], R Kalendar[2], R Ford[3], J Macas[4] and M Griga[1]
[1]*Agritec Plant Research Ltd, Plant Biotechnology Department, Šumperk, Czech Republic;* [2]*Institute of Biotechnology, MTT/BI Plant Genomics Lab, Helsinki, Finland;* [3]*BioMarka, Melbourne School of Land and Environment, The University of Melbourne, Victoria, Australia and* [4]*Biology Centre ASCR, Institute of Plant Molecular Biology, České Budějovice, Czech Republic*

A detailed examination of 45 pea (*Pisum sativum* L.) simple sequence repeat (SSR) loci revealed that 21 of them included homologous sequences corresponding to the long terminal repeat (LTR) of a novel retrotransposon. Further investigation, including full-length sequencing, led to its classification as an RLC-*Angela*-family-FJ434420 element. The LTR contained a variable region ranging from a simple TC repeat $(TC)_{11}$ to more complex repeats of TC/CA, $(TC)_{12-30}$, $(CA)_{18-22}$ and was up to 146 bp in length. These elements are the most abundant *Ty1/copia* retrotransposons identified in the pea genome and also occur in other legume species. It is interesting that analysis of 63 LTR-derived sequences originating from 30 legume species showed high phylogenetic conservation in their sequence, including the position of the variable SSR region. This extraordinary conservancy led us to the proposition of a new lineage, named *MARTIANS*, within the *Angela* family. Similar LTR structures and partial sequence similarities were detected in more distant members of this *Angela* family, the barley BARE-1 and rice RIRE-1 elements. Comparison of the LTR sequences from pea and *Medicago truncatula* elements indicated that microsatellites arise through the expansion of a pre-existing repeat motif. Thus, the presence of an SSR region within the LTR seems to be a typical feature of this *MARTIANS* lineage, and the evidence gathered from a wide range of species suggests that these elements may facilitate amplification and genome-wide dispersal of associated SSR sequences. The implications of this finding regarding the evolution of SSRs within the genome, as well as their utilization as molecular markers, are discussed.
Heredity (2009) **103,** 157–167; doi:10.1038/hdy.2009.45; published online 22 April 2009

Keywords: legumes; microsatellite; pea; phylogeny; retrotransposon; repeat variation

Introduction

Microsatellites or simple sequence repeats (SSRs) are composed of tandemly arranged repetitions of identical, 1- to 5-bp motif-forming arrays, which are usually tens to hundreds of nucleotides long (Tautz and Renz, 1984). Microsatellites are abundant within both animal and plant genomes (Temnykh *et al*., 2001; Morgante *et al*., 2002; Mun *et al*., 2006) and exhibit a high mutation rate, making them attractive as multiallelic, codominant and reliably scored molecular markers. In the most commonly used approach, sequence information of microsatellite-flanking regions is employed to design locus-specific PCR primer pairs to identify amplicon size polymorphism among genotypes. Despite the ubiquitous use of SSRs as genetic markers and their impact on medical and biological science (Rubinsztein, 1999), relatively little is understood about their origin and evolution (Schlötterer, 2000). It is assumed that short proto-microsatellites arise at random and the repeats are subsequently extended by a slippage mechanism (Schlötterer, 2000; Buschiazzo and Gemmell, 2006). This, together with low selection pressure, is proposed to result in high allelic diversity at the microsatellite loci. The origin from corresponding proto-microsatellite sequences may be rapid, as judged from the comparison of SSR loci between two rice subspecies (*Oryza sativa* subsp. *indica* and *japonica*) (Gao and Xu, 2008), when microsatellite motifs were often absent in one of the genomes. Comparative studies among several plant species, with a 50-fold range in genome size, revealed a positive relationship between genome size and the proportion of repetitive DNA (Morgante *et al*., 2002). It is interesting that SSRs were thought to be preferentially associated with the low-copy, gene-rich regions of *Arabidopsis*, corn, rice, soybean and wheat (Morgante *et al*., 2002), supported by a recent study of the model legume *Medicago truncatula* (Mun *et al*., 2006). These findings are reflected in the relatively high frequency of SSR transferability among legume species (Pandian *et al*., 2000; Gutierrez *et al*., 2005). The contradictory view that microsatellites are derived from repetitive (high copy number) sequences is supported by several reports in plants (Ramsay *et al*., 1999; Temnykh *et al*., 2001; Koike

Correspondence: Dr P Smýkal, Agritec Plant Research Ltd, Plant Biotechnology Department, Zemědělská 2520/16, CZ-787 01 Šumperk, Czech Republic.
E-mail: smykal@agritec.cz
Received 19 November 2008; revised 11 March 2009; accepted 19 March 2009; published online 22 April 2009

et al., 2006; Tero *et al*., 2006), insects (Wilder and Hollocher, 2001; Meglecz *et al*., 2007; Van't Hof *et al*., 2007) and nematodes (Johnson *et al*., 2006), suggesting a link between transposable elements and microsatellite sequence. One theory is that the transposable element harboring the proto-microsatellite sequence distributes it genome-wide by transposition and this sequence subsequently develops into a full microsatellite. Thus, a large number of microsatellites may be generated from one identical sequence with an intrinsic property to mutate.

In pea (*Pisum sativum* L.), several marker systems on the basis of either SSR (Loridon *et al*., 2005; Smýkal *et al*., 2008) or retrotransposons (Flavell *et al*., 2003; Kalendar and Schulman, 2006; Smýkal, 2006) have been developed and used for germplasm characterization. In this work, we present evidence that large parts of these SSR loci are derived from microsatellite motifs embedded within long terminal repeats (LTRs) of a retrotransposon belonging to the *Angela* family of the *Ty1/copia* superfamily. These elements are abundant in the pea genome and were found to be conserved across various legume taxa. Moreover, the LTR structure was found to be conserved for the whole *Angela*-family retrotransposons, including elements from a wide range of dicot and monocot species. Thus, it is hypothesized that these elements contribute substantially to generation and dispersion of microsatellite repeats in angiosperm genomes.

Materials and methods

Plant material

Fabaceae genera and species: *Pisum abyssinicum*, *P. fulvum*, *P. sativum* ssp. *elatius*, *P. sativum* ssp. *sativum* cv. Bohatýr, ssp. *arvense*, ssp. *hortense* (and varieties), *Vicia faba* (and varieties), *Lupinus angustifolius*, *Cicer arietinum*, *Glycine max*, *Phaseolus vulgaris*, *Trifolium repens*, *T. pratense*, *Lens culinaris* (AGRITEC Ltd legume collection, Šumperk, Czech Republic), *Lotus corniculatus*, *Medicago sativa*, *M. truncatula*, *Lathyrus sativus*, *L. nissolia*, *L. sativus*, *L. pratensis*, *L. vernus*, *L. tinigitanus*, *L. sylvestris*, *L. tuberosus*, *L. hirsutus*, *Vicia sativa*, *V. melanops*, *V. pisiformis*, *V. onobrychoides*, *V. michauxi*, *V. narbonensis*, *Arachis hypogea* (Crop Research Institute, Prague, Czech Republic), *Vigna radiata* (market in India), *Gleditchia triacanthos*, *Robinia pseudoaccacia*, *Wistaria chinensis* (Botanic Garden, Palacky University, Olomouc, Czech Republic), tropical legume trees: *Acacia karroo*, *Bauhinia sp.*, *Cassia sp.*, *Delonia regia*, *Erythrina caffra*, *Leucena leucephala*, *Sophora tomentosa*, *Senna meridionalis*, *Schotia afra*, (Prague Botanic Garden, Prague, Czech Republic), *Calopogonium mucunoides*, *Cajanus cajan*, *Lablab purpureus*, *Mucuna prariensis*, *Stylosanthes capitata*, *Macrotylona axillara* (Institute of Field and Vegetable Crops, Novi Sad, Serbia).

DNA isolation

Young leaves from greenhouse grown plants were harvested and stored at $-80\,^{\circ}C$ until DNA isolation. Genomic DNA was isolated using the Invitek plant genomics DNA column protocol (Invitek, Berlin, Germany). DNA obtained from approximately 100 mg of fresh weight leaf material was resuspended in TE buffer at $\sim$50–100 ng μl^{-1} and stored at $-20\,^{\circ}C$. DNA quality was assessed electrophoretically and spectrophotometrically.

Pea SSRs amplification

A total of 45 selected distinct pea SSR loci (A-5, A-6, A-9, B-11, B-14, C-20, AA-5, AA-31, AA-37, AA-67, AA-90, AA-92, AA-94, AA-98, AA-99, AA-122, AA-195, AA-205, AA-206, AA-224, AA-238, AA-278, AA-285, AA-315, AA-317, AA-321, AA-355, AA-387, AA-398, AA-473, AB-23, AB-29, AB-44, AB-65, AB-69, AB-81, AB-122, AB-130, AB-184, AC-58, AD-131, AD-146, AD-147, AD-175, AD-237, AD-270) were amplified from *P. sativum* cv. Bohatýr, using the conditions and primers of Loridon *et al*. (2005).

Cosmid screening and analysis

To identify full-length elements, a pea genomic library, prepared using a cosmid vector (Neumann *et al*., 2003), was screened (3456 clones) using an LTR probe prepared from the sequenced pea A-9 SSR clone with PCR primers LTR-3F and LTR-4R (Table 1, Figure 1). A total of 40

Table 1 List of primers used in the study

Denotation	*Orientation*	*Primer position*	*Sequence 5′–3′*
LTR region		*Referred to c4 LTR*	
LTR-1F	Forward	1–23	TGTTGGTGTAAGCCCCTAGAGGCC
LTR-2F	Forward	1046–1072	GTGGTTCTATAAATAGAACCCTTGTG
LTR-3F	Forward	1185–1207	GTTCGTGTGGACTGAGTAGAGA
LTR-1R	Reverse	1–23	GGCCTCTAGGGCTTACACCAACA
LTR-2R	Reverse	1185–1207	TCTCTACTCAGTCCACACGAAC
LTR-3R	Reverse	1353–1381	GTGATCCTTACGAAGGGGCATGATCAGTG
LTR-Ps	Reverse	1407–1432	TGAAGGAGAATTGCGACCCAAAGCGC
LTR-Mt	Reverse	1375–1396	TGAAGGGATTTTAAGAGGGGT
Internal region		*Referred to c4 internal sequence*	
F529	Forward	374–393	GTCAAGAGTTATACATTAG
F1129	Forward	975–997	GAGAAGCTTAGATCGTTATTAG
F1810	Forward	1635–1656	CATGTAGATTACCATGACAAC
F2437	Forward	2263–2283	AAGTCCATTCTAATGATCGG
R2490	Reverse	3156–3177	CCTCCTCTGGCTTGTATGTTC
R1795	Reverse	3850–3870	GTGTATCGATGCTTTGTGAT
R1157	Reverse	4489–4508	GCTTCTGGAAGCTTGCTTC
R498	Reverse	5128–5148	AGCTAGCATCGGTGTATCCA
PPT	Forward	5530–5557	GACAAAGCCTCTTGCGCAGCAGAAGCA
PBS	Reverse	46–73	TAAACAACTATCAGATGCATGGTTTCG

Abbreviations: LTR, long terminal repeat; PPT, polypurine tract; PBS, primer-binding site.

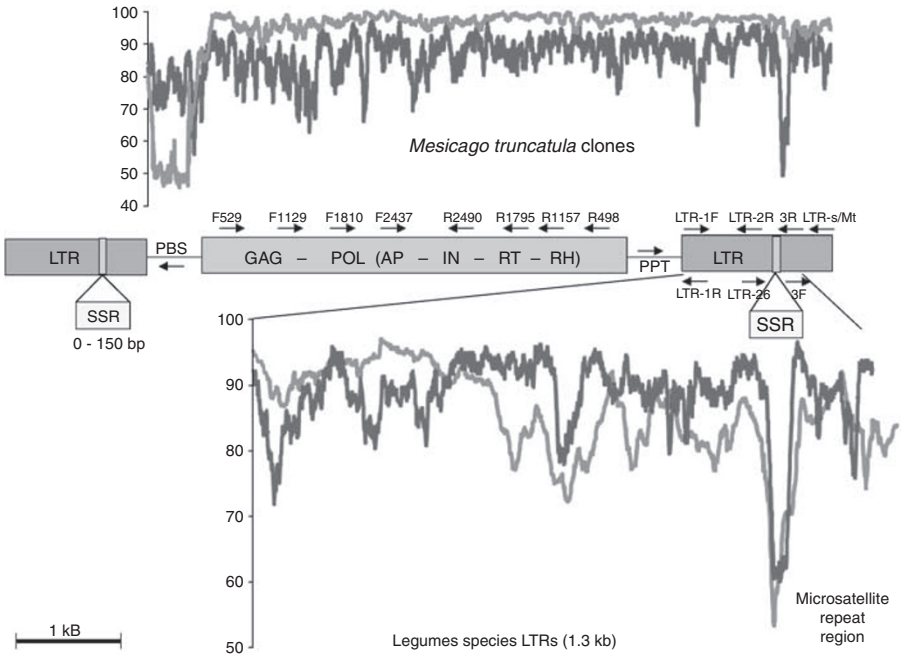

Figure 1 Schematic representation of the 8.45 kb RLC-*Angela*-FJ434420 retroelement with indicated microsatellite sites, with computed similarity in red and linguistic complexity in blue, calculated as percentage values over a 50 bp window from 11 full-length *Medicago truncatula* homologues (above) and the 1.3 kb LTR region of 63 sequences from 30 legume species (below). Long terminal repeat (LTR) region harboring simple sequence repeat (SSR), primer-binding site (PBS), polypurine tract (PPT) and protein domains of gag-pol, putative capsid protein (GAG), aspartic proteinase (AP), IN—integrase, reverse transcriptase (RT), RNAseH (RH), are shown. A full colour version of this figure is available at the *Heredity* journal online.

cosmid clones were selected (from 480 positives) after replicated filter screening. The cosmid DNA was subjected to enzyme digestion with *Hind*III, *Bam*HI or *Eco*RI using the manufacturer's protocol (Fermentas, Prague, Czech Republic), followed by electrophoresis on 1% agarose gel. Fragments were blotted onto a nylon membrane and hybridized with a PCR DIG-labeled (Roche, Mannheim, Germany) LTR probe (outside the SSR repeat) (Table 1) to discriminate between solo- and multiple-LTR clones (likely harboring full-length elements). Only those clones providing signal on at least two different bands (LTR fragments) were used for further analysis. Corresponding bands were excised and cloned into the pSK Bluescript vector (Stratagene, La Jolla, CA, USA).

Legume LTR amplification

To retrieve partial LTR sequence from investigated legume species, LTR sequence deduced from sequenced pea SSR loci as well as the full-length element were used to design primers annealing to various parts of the predicted LTR region (Table 1). The LTR-1F and LTR-3R primers amplified the 5′ proximal (1.0–1.3 kb) LTR region. For microsatellite repeat length and composition investigation, the LTR-2F and LTR-3R primers were used to amplify a 300- to 400-bp fragment. To test for the presence of variable 3′ LTR end sequences, primers homologous to either the earlier identified pea c732 sequence (LTR-Ps) or retrieved from the *Medicago* database (LTR-Mt) were used together with the LTR-1F primer (Table 1, Figure 1).

Cloning and sequencing

Clear and reproducibly amplified bands, from either pea SSR loci or legume species LTR fragments, were excised from the agarose gel, purified using the QIAquick gel extraction kit (Qiagen, Hilden, Germany) and cloned into the pJET vector using the JET cloning kit (Fermentas). Transformed *E. coli* colonies were picked from plates and screened for the presence and approximate size of inserts using the vector-derived PCR primers. The plasmid DNA was extracted using the JET miniprep kit (Fermentas) and the insert was sequenced from both ends on an ABI 310 automated sequencer using the ABI PRISM BigDye Terminator Cycle Sequencing kit (Applied Biosystems, Foster City, CA, USA).

IRAP analysis

To estimate genomic abundance, inter-retrotransposon amplified polymorphism (IRAP) analysis was carried out according to the method of Smýkal (2006), adapted from Kalendar and Schulman (2006) with the LTR-1R and LTR-3F primers (Table 1). All PCR reactions were carried out as described in Smýkal (2006) and Smýkal *et al*. (2008).

Sequence assembly and analysis

CLUSTALW alignment was carried out at http://pbil.ibcp.fr. The final alignment was visualized with Boxshade 3.21 (http://www.ch.embnet.org/software/BOX_form.html) or MEGA 4.0 software (http://www.megasoftware.net). Sequence assembly, primer design and restriction analyses were carried out with FastPCR Professional software version 5.1.83 (Primer Digital Ltd, Helsinki, Finland). Conserved protein domains were searched with RPS-Blast. Sequences of tRNAs were used for identification of the primer-binding site from the *A. thaliana* tRNA database (Lowe and Eddy,

1997). Nucleotide diversity was determined using DnaSP 4.10 (http://www.ub.es/dnasp/interface.html). For estimation of the insertion time of the retroelement, pairwise comparison of two corresponding LTR regions was carried out with MEGA 4.0 software (http://www.megasoftware.net) using Kimura's 2 parameter model (http://www.megasoftware.net) using the formula: $T = K/2r$, where T is time of insertion, K is divergence parameter and r is average substitution rate (taken as Ks value). The molecular clock was calibrated with Ks values of synonymous nucleotide substitution rate calculated by Jing *et al*., (2005), with 7.0×10^{-9} substitutions per site per year.

Database searches and similarity analysis

BLAST (Altschul *et al*., 1990) analysis was carried out on NCBI (http://www.ncbi.nlm.nih.gov) and GENOME (http://www.genome.jp) servers using default settings. Dot-plot sequence comparisons were carried out with the Gepard program (http://mips.gsf.de/services/analysis/gepard). The following specific legume databases were used to identify homologs: Soybase (http://soybase.org/), the *Medicago* Genome database (http://www.medicago.org), http://www.medicago.org, the *Lotus japonicus* Kazuza database (http://www.kazusa.or.jp/lotus/) and the BeanGenes database (http://beangenes.cws.ndsu.nodak.edu/). In addition, databases for *Triticeae* and *Avena* (http://wheat.pw.usda.gov/GG2/genomics.shtml) and the RetroOryza database (http://www.retroryza.org) were also searched. To identify the position of *M. truncatula* elements on chromosomes, CViT BLAST (http://medicago.org/genome/cvit_blast.php) was used.

Linguistic complexity, similarity and CG-skew analysis

The sequence analysis complexity calculation method was used to search for conserved regions in comparative sequences, for the detection of low-complexity regions including SSRs or imperfect direct or inverted repeats. The linguistic complexity measurement was carried out using the alphabet-capacity l-gram method (Orlov and Potapov, 2004) within a sliding window of 50 bp with one base assessed at a time. The profile was constructed by averaging the complexity values along all sequences of a set in each window. The complexity values were converted to a percentage value, in which 100% means maximal 'vocabulary richness' of a sequence. The DNA sequence is studied as text in the four-letter alphabet; the repetitiveness of such a text is correlated with repetition of some k-grams (words) and served as a measure of sequence complexity. The more complex a DNA sequence is, the richer is its oligonucleotide vocabulary. Similarity analysis was carried out with sequences from multiple alignments and was calculated in sliding windows with length of 50 nucleotides. The similarity values were converted to a percentage, in which 90% similarity mean identity at 45 out of 50 bases in the window, with 100% being the highest level of sequence similarity. The GC skew in a sliding window of 100 bp was calculated with a step of one base, according to the formula, GC skew = (G–C)/(G + C), in which G is the total number of guanines and C is the total number of cytosines for all sequences in the windows (Schneeberger *et al*., 2005). Positive GC-skew values indicated an overabundance of G bases, whereas negative GC-skew values represented an overabundance of C bases.

Southern hybridization

Southern hybridization was carried out to verify, independently of PCR, the presence and abundance of the retrotransposon in the investigated species. Fragments for Southern hybridization were amplified using with LTR-1F and LTR-3R primers (1000 bp) in case of LTR or with F1810 and R1157 primers in case of gag-pol probes using the above mentioned PCR conditions (Table 1, Figure 1) and the sequenced pea cosmid c4 clone used as the template. These amplicons were labeled using the DIG PCR labeling mix (Roche). A total of 30μg of target species genomic DNA was restricted with *Eco*RI (Fermentas) using the manufacturer's protocol, and the fragments were resolved on 1% agarose gel. Southern blotting was carried out by capillary transfer onto a nylon membrane (Roche) with heat fixation. Hybridization was carried out according to the manufacturer's manual using the DIG EasyHyb kit (Roche) solution with final high-stringency washes of $0.2\times$ SSC (sodium chloride–sodium citrate solution) and 0.1% sodium dodecyl sulfate at 65 °C for 2×20 min. A chemiluminiscent signal was recorded on Medix-XBU X-ray film (Foma, Hradec Králové, Czech Republic).

Results

Variable SSR repeats identified within the LTR part of a pea *Ty1/copia* retrotransposon

Following sequence analysis of 45 pea SSR loci (Loridon *et al*., 2005) used for pea germplasm characterization (Smýkal *et al*., 2008), 21 contained similar sequences surrounding the variable SSR motif. The similarity included most of the amplified loci except at the 3′ end, in which it abruptly terminated at a conserved CCTTCA motif, which preceded a divergent sequence that was different for each locus (Figure 2). Such characteristics suggested that these SSR sequences could be a part of a retrotransposon LTR region and that the loci shown on Figure 2 represented various insertion sites within the pea genome.

Comparison of all available sequences revealed considerable diversity of the microsatellite region. This varied from a simple TC repeat $(TC)_{12}$ in the B-11 locus to more complex TC/CA $(TC)_{12-31}$, $(AC)_{10-29}$ and TC/ATCT/CA type repeats that were up to 146 bp in length (AD-175) (Figure 2). Using pea derived locus-specific primers from Loridon *et al*., (2005), all 21 SSR loci were amplifiable from *P. sativum*, *P. abyssinicum*, *P. fulvum* and *P. elatius* species, but not from other legumes, including the *Viceae* tribe genera (data not shown). Moreover, the application of four of the 21 SSR loci (A-9, B-14, AD-175, AA-321) for genetic relationship analysis across 800 *P. sativum* accessions had earlier revealed a high level of length polymorphism and indicated further repeat expansion, even within the cultivated pea genepool (Smýkal *et al*., 2008).

To obtain a genuine full-length genomic sequence of the newly identified element, a pea genomic cosmid library (Neumann *et al*., 2003) was screened with the LTR-specific probe. The proportion of positive clones

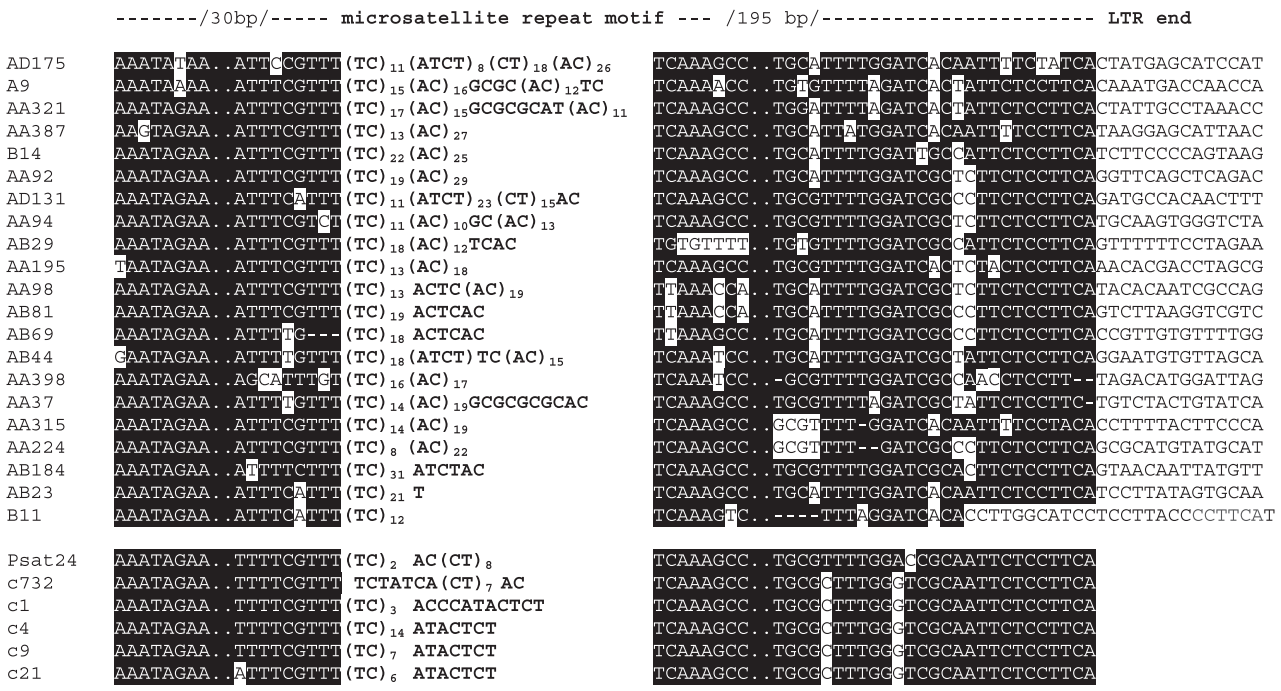

Figure 2 CLUSTALW alignment of the terminal part of pea simple sequence repeat (SSR) loci together with the pea Psat24, c732 sequences and cosmid clones c1, c4, c9 and c21. The ends of the putative long terminal repeat (LTR) region and microsatellite repeat motifs are shown. Number of bases shown above the alignment indicates the length of the regions omitted (as shown by dots) from the alignment.

(480 out of 3456 clones, corresponded to about 17 000 copies per 1C) confirmed an earlier estimate of high abundance of this element in the pea genome (Macas *et al*., 2007). Subsequent sequencing of the selected cosmid clone c4 (FJ434420) yielded an almost complete copy of the element, which was 8423 bp long and was slightly truncated by cloning, missing only 51 bp from the 3′ end. The element was designated as belonging to the *RLC-Angela*-family-FJ434420 (according to Wicker *et al*., 2007) and had structural and functional features typical for LTR-retrotransposons of the *Ty1/copia* superfamily (Figure 1). On the basis of sequence similarities to a number of elements from various taxa, including barley BARE-1, rice RIRE-1 and earlier phylogenetic analysis of Ps-copia-1/175 (Macas *et al*., 2007), the newly identified element was assigned to the *Angela* family. An *in silico* identification on the basis of the assembled Ps-copia-1/175 (Macas *et al*., 2007) was used for revealing the coding region sequence organization. Database searches identified high similarities of this element to a number of legume repetitive sequences, including a 1432-bp solo-LTR within the pea cosmid clone c732 (GenBank accession FJ405426) and a putative retrotransposon within the pea BAC sequence (CU655881).

Phylogenetic conservation of the RLC-*Angela*-family-FJ434420 LTR region in *Fabaceae*

After pea cosmid and SSR sequence isolation and *M. truncatula* homolog retrieval, a search for homologous elements in other legume species was carried out using a PCR approach. As the 3′ end of the *Pisum* and *Medicago* LTR sequences varied widely, primers were designed to anneal to the more highly conserved 5′end. A fragment of the expected 1.2 kb size was amplified in all 49 legume species tested, representing various tribes including tropical trees (data not shown). A total of 63 sequences from 30 legume species were sequenced and aligned, revealing their high similarities (80–90%) to the pea LTR sequence. The most divergent region was between the 1081 and 1123 bp positions (in reference to the LTR of the pea cosmid c4 FJ434420 sequence), in which an SSR was located in all cases. This was evidenced by low similarity and low linguistic complexity over 50-bp window scans (dropping from an average of 90–55% in this region, Figure 1). Variable repeat motifs comprised short $(TC)_2$ repeats in *V. sativa* and *L. sativus*, $(TC)_{19}$, $(TC)_{20}$, $(TC)_{25}$ repeats in *Delonia*, *Lupinus albus*, *T. repens* and *G. max* and more complex $(TC)_5$ ATCT repeats in *V. faba* (Figure 3). Despite high overall sequence homology, the LTR sequence varied within species, particularly between the 20 to 80 bp positions immediately 3′ to the SSR (Figure 1).

To investigate the divergence found between *Pisum* and *Medicago* sequences at the 3′ LTR end, two respective reverse primers were designed (Table 1, Figure 1). When the pea LTR-Ps reverse primer was used together with the LTR-1F primer, a fragment representing the entire LTR sequence was obtained from all *Arachis, Glycine, Pisum, Lathyrus, Lupinus, Lens* and *Vigna* samples (the amplification specificity was confirmed by partial sequencing; data not shown), and only a weak product was amplified from *Cicer, Medicago, Phaseolus* and *Vicia*. No product was obtained from more phylogenetically distant legume species (data not shown), whereas use of primer matching to the *M. truncatula* sequences (LTR-Mt) yielded the expected product from *Medicago, Phaseolus, Vigna, Glycine* and *Trifolium* samples (data not shown), and this product was confirmed by partial sequencing.

Figure 3 Sequence of long terminal repeat (LTR)-localized simple sequence repeats (SSR) motif and adjacent parts of 30 sequences of 30 legume species, taken from CLUSTALW alignment of 1.2–1.4 kb LTR parts. Sequence shown in black letters show unique regions adjacent to SSR and bases boxed in black are conserved in at least half of the sequences. The number of bases indicates the length of the regions omitted from the alignment, in reference to the pea c4 LTR (FJ434420) sequence.

Inter-retrotransposon amplified polymorphism analysis with LTR-derived primers enabled estimation of the abundance of the RLC-*Angela*-family-FJ434420 elements among legume species. A multilocus profile was amplified from all investigated legume species, especially from those belonging to the *Papilionoideae* section. This varied in complexity from numerous bands in *Pisum* to a few distinct bands in *Vicia*, *Trifolium*, *Medicago* and tropical trees species (data not shown). It is interesting that a high level of polymorphism was detected even within genotypes of a single species, as investigated in *Pisum*, *Lupine*, *Vicia* and *Lathyrus* (data not shown). Among cultivated pea varieties, the level of polymorphism was as high as 25 of 38 detected fragments (66%). This indicated that although the homologous LTR sequences may be present, their genomic distribution and abundance might differ significantly from those in other genotypes or species.

The widespread occurrence and conservation of *Angela*-family elements in legumes assessed by PCR-based methods was also confirmed using Southern blotting. The probes derived from parts of the LTR or gag-pol regions generated distinct signals on genomic DNA samples from *P. sativum*, *M. sativa*, *L. sativus*, *T. repens*, *L. culinaris* and *V. faba*, and weaker, but still clearly detectable, signals in *Phaseolus*, *Lupinus*, *Arachis*, *Cicer* and *Glycine* samples (data not shown).

Owing to its extraordinarily conserved nature, we propose to separate this distinct lineage of legume *Angela*-family retrotransposons, and name it as the *MARTIANS*.

Comparative analysis of *Angela*-family elements from various taxa

Availability of one (c4) full-length (RLC-*Angela*-family-FJ434420) element, together with LTR pairs of partially sequenced (c1, c9, c21) clones isolated from a cosmid pea library, enabled testing of its antiquity. Pairwise nucleotide sequence alignment and comparison of the c1, c4, c9 and c21 LTR pairs identified 6, 25, 8 and 14 nucleotide substitutions, 1, 7, 0 and 9 single bp insertions and 0, 11, 0 and 0 bp long deletions, respectively. The calculated nucleotide diversity among LTRs within clones was 0.0093(c1), 0.0134 (c4), 0.0065 (c9) and 0.0254 (c21), (0.0136 on average), the ratio of nucleotide transitions to transversions was 1.4, 2.3, 0.3 and 3.1 (1.4 on average). Subsequently, calculation of antiquity suggested an age range from 2 to 5 Mya. All clones contained a TC repeat motif located in both LTR regions at an identical position of 1006 bp from the 5′ end of the LTR. The c1 clone contained a $(TC)_3ACC(TC)_5$ in both LTRs, the c4 contained $(TC)_{14\ and\ 13}$, the c9 clone contained an identical $(TC)_{12}$ and the c21 clone contained $(TC)CC(TC)_{4\ and\ 6}$ repeats in the respective 5′ and 3′ LTRs.

Taking advantage of the extensively sequenced genome of *M. truncatula* (*Mt*), we identified 11 full-length and four truncated homologs of *Angela*-family elements from this specie distributed on chromosomes 2, 3, 4, 5, 6 and 7 as supported by a CViT BLAST search. In addition, 44 solo-LTRs were also retrieved (queried in October 2008). It is interesting that all *Medicago* LTRs contained identical, non-expanded putative proto-microsatellite sequence motifs (for example, TATAA, CACA and CTCTCT) and a short variable poly-A region between positions 1030 to 1050 bp. Linguistic complexity analysis of both complete elements and solo-LTRs within the 1.4-kb LTR regions revealed high similarity (80–100%) with the exception of two regions from 50 to 70 bp and 850 to 1000 bp, in which similarity dropped to 65 and 50%, respectively, because of the presence of insertions. Similarly, analysis of the 5.5-kb internal coding region showed high similarity (60–100%) with an exception of the first 400 bp after the conserved primer-binding site (Figure 1). The calculated nucleotide diversity among 11 LTRs ranged from 0.0014 to 0.0285 (average 0.0123) and insertion time was estimated to range from as little as a few thousand up to 3 million years ago. The ratio of nucleotide transitions to transversions ranged from 0.0014 (two substitutions) to 0.0285 (1.4 on average). In contrast to a completely preserved polypurine tract site, the primer-binding site for reverse transcription was variable. Translation of the coding region showed that all

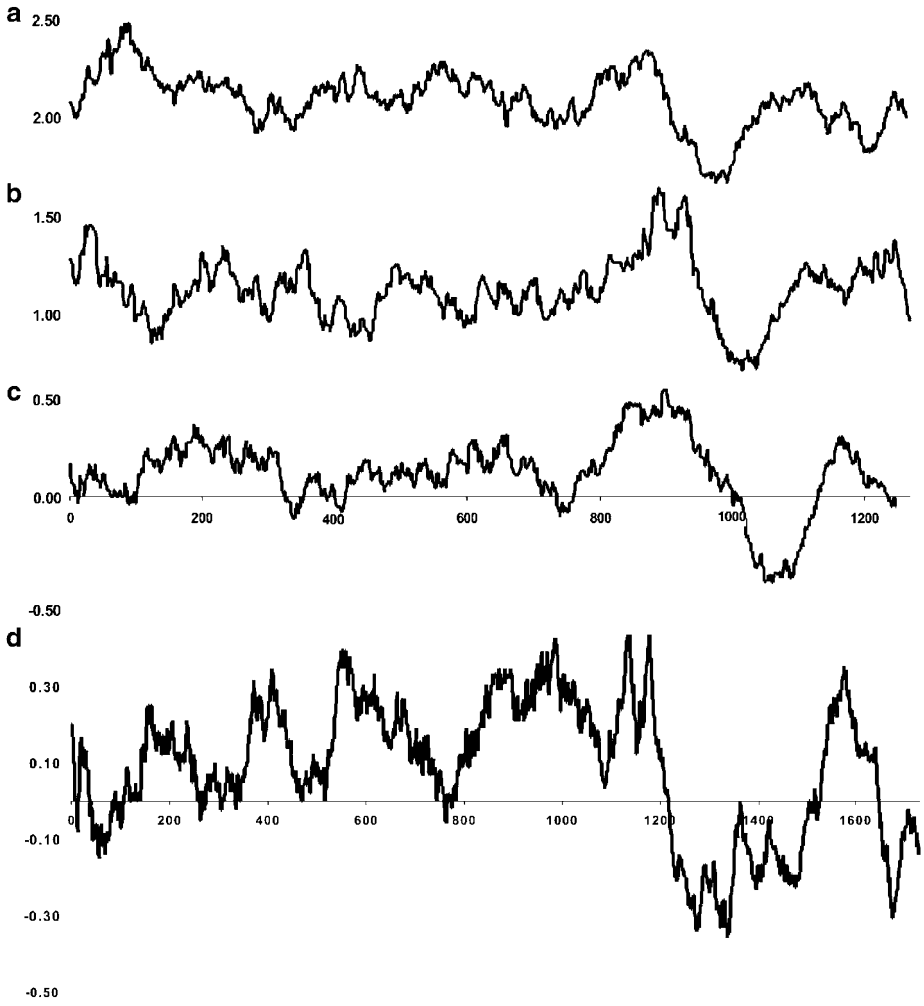

Figure 4 Result of CG-skew analysis of LTR regions, showing the DNA composition computed in a sliding window of 100 bp with a step of one base. (**a**) *Pisum sativum* RLC-*Angela*-FJ434420, (**b**) *Medicago truncatula* homolog, (**c**) *Oryza sativa* RIRE-1, (**d**) *Horderum vulgare* BARE-1. Positive GC-skew values indicate an overrepresentation of G bases, whereas negative GC-skew values represent an overabundance of C bases.

retrotransposons were likely not functional because of mutations accumulation, apart from the reverse transcriptase region, which was identical and intact in three clones. The 5-bp target site duplication was identified, with high preference for thymidine at positions $-1/+1$ (in 37 of the 55 cases) and $-4/+4$ (in 36 of the 55 cases). The presence of an intact target site duplication in the solo-LTRs suggests that intra-element recombination was involved in the removal of the internal part. Truncation within seven of the solo-LTRs seemed to be consistently at the 3′ end. Analysis of the solo-LTR insertion targets showed that 21 clones (47%) resided in other repetitive DNA sequences, 15 (34%) in intergenic and intron regions and only eight (18%) were in the predicted protein-coding sequences.

Analysis of 25 barley BARE-1 and 36 rice RIRE-1 full-length LTR sequences revealed several similar low-complexity regions, usually at the 3′ end of the LTR, which carry pyrimidine or purine-rich sequences, potentially serving as proto-microsatellites. A highly variable pyrimidine-rich (mostly C-rich) sequence, in both BARE-1 and Wis LTRs, was detected immediately preceding a polymerase II TATA promoter region, and a long purine-rich (WIS LTR, 1304–1311 bp; BARE-1 LTR, 1340–1347 bp) region was detected a further 100 bp upstream (Supplementary Figure 1). Both of these regions were highly variable, whereas the sequence between the TATA promoter and the 3′end of the LTR was highly conserved. In contrast, no such variable sequences were detected in RIRE-1 from *O. sativa* in the corresponding region (Supplementary Figure 1). To examine the role of DNA sequence composition of conserved LTR region, we carried out CG-skew analysis, which measures DNA strand composition asymmetry of cytosines compared with guanines on one strand of the double helix. This analysis on LTRs of barley BARE-1 and RIRE-1 from rice showed the same pattern of GC skew (Figure 4c and d). BARE-1 LTR has two TATA boxes promoters at 983 and 1355 bases; these coordinate at opposite side to GC skew and these sites are very different from whole LTR. The maximum GC-skew values overlap well with the TATA-box 1 and symmetrical dramatic decrease in GC skew just downstream overlaps well with second TATA-box 2 promoter.

The symmetry between TATA-box 1 and TATA-box 2 promoters indicated that bases composition around the SSR site is similar for all studied *Angela*-family LTRs from different species (Figure 4). The same places at pea RLC-*Angela*-family-FJ434420 can correspond to region near 850 and 1050 bases (Figure 4a). CT-rich SSR-like region in pea is located at 1095–1141 bases, which coordinates for GC-skew gap, after which the second promoters might be located.

Discussion

Identification of the evolutionary conserved *Angela* retrotransposon family

Long terminal repeats retrotransposons are ubiquitous components of all eukaryotic genomes investigated so far. In higher plants, they often constitute the majority of genomic repeats and their differential amplification contributes significantly to the profound differences in genome size observed between various taxa (Piegu *et al*., 2006). In addition, they have an impact on the genome structure and function as a source of regulatory and coding sequences (Kazazian, 2004). However, despite their ubiquity, after host speciation, retrotransposons have undergone diversification resulting in low sequence homology within genera (Wicker and Keller, 2007).

The initial analysis of pea microsatellite loci-flanking regions in this study indicated an association with the LTR region of a novel retrotransposon. Sequencing of the entire element identified an 8.5 kb large *Ty1/copia* retrotransposon of the *Angela* family, homologous to the recently *in silico*-identified Ps-copia-1/751 element (Macas *et al*., 2007). The level of conservation to distantly related legume species is thought to be unique among described retrotransposons. This homology is even more pronounced when considering the similarity detected in rapidly evolving LTR regions (Vicient *et al*., 2005). This is in contrast to the highly conserved coding regions, such as the reverse transcriptase, used for family definition (Wicker *et al*., 2007). Relatively few studies have investigated LTR sequence conservation across species (Vicient *et al*., 1999, 2005), and furthermore no sequence similarity has been reported in the legumes. Thus, high conservation detected in the RLC-*Angela*-family-FJ434420 element in this study, which shows extensive (over 80%) similarity in the LTR regions across the large *Fabaceae* family, is unique. Until now, such homologs were only detectable across the closely related genera and to limited extent within the *Triticeae* tribe using the barley BARE-1 or rice RIRE-1 sequences (Vicient *et al*., 1999, 2005; Piegu *et al*., 2006; Roulin *et al*., 2008). This suggests that sequence conservation may be a common feature of this lineage of *Angela* family, which was shown to be one of the six ancient lineages (families) of retrotransposons that existed before the divergence of monocots and dicots, 150 My ago (Wicker and Keller, 2007).

The authors have hypothesized that specific selection forces may act on *copia* elements. Horizontal transfer across species boundaries, leading to homogenization of the retrotransposon gene pool, was also proposed as the most likely explanation of conservation (Wicker and Keller, 2007). Indeed, the multiple horizontal transfer of rice RIRE-1 to seven *Oryza* species was recently reported (Roulin *et al*., 2008). The opposing notion that retrotransposons have undergone stabilizing selection since the monocot–dicot divergence is currently considered unlikely, although there is an evidence of stabilizing selection on animal elements (Katzman *et al*., 2007).

It is notable that the RLC-*Angela*-family-FJ434420 belongs to the same family of highly abundant elements as barley BARE-1 (Vicient *et al*., 1999), wild rice RIRE-1 (Piegu *et al*., 2006) and wheat WIS and *Angela* (Vitte and Bennetzen, 2006), all with over 100 000 copies estimated per respective genome. The RLC-*Angela*-family-FJ434420 was also determined to be the most abundant *copia* type element in the pea genome (Macas *et al*., 2007), with an estimated 8000 full-length and 9000 solo-LTR copies per 1C (comprising 2% of the pea genome). Similarly, the rice RIRE-1 element (Noma *et al*., 1997) contributed substantially to expansion of the wild-rice genome (965 Mbp of *O. australiensis* versus 390 Mbp of *O. sativa*), and comprised about 250 Mbp (26%) of the genome (Piegu *et al*., 2006).

Retrotransposons as microsatellite dispersal agents

In this study, the LTR retrotransposons were shown to contribute to the origin and distribution of microsatellite repeats. Such information until now has not been readily available because of the limited studies that have been conducted on retrotransposon LTR regions, despite their functional importance. Certain proto-microsatellite sequences present within the LTR of the *Angela* family of *Ty1/copia* retrotransposons in a broad range of both mono- and dicotyledonous species are thought to have been spread throughout the genome by transposition activity, in which they subsequently expanded into variable size and composition microsatellite loci. Thus, a large number of proto-microsatellites could be generated from one original sequence with an intrinsic property to develop into a microsatellite. Although such association between mobile elements and microsatellite repeats has been documented (Ramsay *et al*., 1999), the actual molecular mechanisms underlying this relationship are not well understood. This is intriguing, especially in relation to the occurrence of SSRs, which occur far more often than by chance in the genomes of all eukaryotes (Temnykh *et al*., 2001; Gao and Xu, 2008). The reasons for this are not yet fully understood, but there is an evidence that microsatellites function in gene regulation (Zhang *et al*., 2006) and meiotic recombination (Bagshaw *et al*., 2008).

Also of interest is the detection of variable SSR repeats at specific sites of LTR regions in all the analyzed legume sequences presented in this work, including four sequenced pea cosmid clones, one pea BAC and partial clover BAC sequences. This together with the identified pea SSR loci suggests that indeed the RLC-*Angela*-family-FJ434420 element and its possible homologs carry and disperse microsatellite repeats in the pea and possibly other legume genomes.

Such SSR association with a retrotransposon is in agreement with the data of Temnykh *et al*. (2001) and Inukai (2004) on rice, of Koike *et al*., (2006) on wheat and barley (Ramsay *et al*., 1999; Vicient *et al*., 2005) as well as animals (Wilder and Hollocher, 2001; Meglecz *et al*., 2007; Van't Hof *et al*., 2007). In particular, the study of Ramsay *et al*. (1999) showed an intimate association of SSRs with

retrotransposons in barley. These authors identified repeats with homologies to three different high-copy retrotransposons, such as BARE-1 /WIS2-1A, rye R173 and millet PREM1. Although varied in length and composition, AC, AG and remarkably CT repeat types prevailed. In contrast to this study, with repeat region located within the LTR, in the study of Ramsay *et al*. (1999) barley SSRs were located in regions adjacent to LTRs (in the vicinity of 5–50 bp).

Plant LTR retrotransposons are thought to be expressed by a polII promoter mechanism, thus being fully dependent on the host. The expression is driven from the 5′ LTR, initiated 3′ to the TATA box, and extends until the 3′R region within the 3′ LTR (Suoniemi *et al*., 1996). The same authors suggested that new promoter sequences that arose by variation may permit the expression of the retroelement in various conditions. Vicient *et al*., (2005) showed that the most variable region of BARE-1 was the U3 region of the LTR, immediately upstream of the putative TATA box. On the basis of the analysis in the current study, it may be speculated that the repeat region before the TATA box is involved in the formation of H-DNA structures. Thus, CT- or A-rich regions may be involved in transcription initiation (Lu *et al*., 2003) and/or help to form RNA secondary structures (Han and de Lanerolle, 2008). Similarly, a comprehensive investigation of T-DNA integration sites in *Arabidopsis* by CG/AT-skew profiles has shown correlation with DNA sequence composition (Schneeberger *et al*., 2005), affecting DNA bending. Using the identical approach, we have found identical sequence composition of LTR regions among barley BARE-1, rice RIRE-1 and this study identified *Medicago* and pea RLC-*Angela*-family-FJ434420 retrotransposons. As in case of BARE-1 LTR, two TATA boxes were identified (Vicient *et al*., 2005), which coordinate at opposite side to GC skew, we hypothesize homologous regions to have identical functions in LTR regions of other species.

It is noted that a recent comprehensive analysis of 24 plant species, including mosses, fern, conifers and monocots/dicots, showed that AG/CT and AC/GT are by far the most abundant and length-variable of all repeats (von Stackelberg *et al*., 2008).

Moreover, the current study brings a novel view in terms of both SSR repeat and LTR sequence conservation. Such conservation in wide range of species is suggestive of selection. This feature is rather unique, as retrotransposons, apart from the evolutionarily conserved reverse transcriptase, are rather diverse outside the species level. On the basis of specific site repeat localization, our data indicated that proto-microsatellite sequences were likely to have been present in the LTR region before diversification of the *Fabaceae*. Such findings are supported by the study of microsatellite genesis across several species in *Diptera* (Wilder and Hollocher, 2001), *Lepidoptera* and *Coleoptera* (Van't Hof *et al*., 2007).

The observation of variable repeat structure, embedded at identical positions in highly homologous flanking sequence of LTR in pea, suggests a single proto-microsatellite origin, which provides a unique opportunity to study SSR changes and may provide valuable information on microsatellite repeat genesis. Slipped strand mispairing during DNA replication is currently thought to cause most microsatellite mutations, but it has also been proposed that unequal meiotic recombination could drive microsatellite evolution (Schlötterer, 2000). In our study, the LTR-localized SSR loci showed signs of repeat turnover, for example, change of motif, in which an original CT motif developed into an ATCT and/or CA in the final composite repeat, as found in *Silene* (Tero *et al*., 2006). In the pea cosmid clones assessed, differences in SSR composition between respective LTR pairs indicated that a slippage mechanism was involved in microsatellite repeat genesis. In retrieved *M. truncatula*, as well as barley and rice homologs, only short proto-microsatellite variable repeat sequences were found in respective LTR regions. In accordance with the findings of Temnykh *et al*. (2001) and Ramsay *et al*. (1999), it may be hypothesized that a substantial proportion of the 340 *Pisum* SSRs currently available (Agrogene Pisum SSR consortium, Loridon *et al*., 2005) are embedded in retroelements. This is likely to be the reason for the frequent failure to map these SSRs as reported by Loridon *et al*. (2005).

Although the presence of an SSR motif within an abundant repeat may impede marker development, the association between SSRs and mobile elements has facilitated the development of an informative retrotransposon-microsatellite amplified polymorphism marker system (Kalendar and Schulman, 2006). This feature, together with their abundance and conserved nature, indicates that the *MARTIANS* lineage of retrotransposons may provide informative, portable markers for a wide range of legumes, the third largest plant family with over 18 000 species.

Supplementary material

The nucleotide sequence data reported in this paper will appear in the EMBL, GenBank and DDBJ nucleotide sequence databases under the following accession numbers. Pea SSR loci: FJ434421-FJ434441, legume LTRs: FJ434442-FJ434452 and FJ409873-FJ409890, full-length pea c4 cosmid clone FJ434420, solo-LTR of pea cosmid clone c732: FJ405426.

Acknowledgements

This work was supported by the Ministry of Education of the Czech Republic (MSM 2678424601 and LC06004 projects), and by the project AV0Z50510513 from the Academy of Sciences of the Czech Republic. The excellent technical support of Mrs E Fialová is greatly acknowledged. PS acknowledges the OECD Fellowship during which, in the highly stimulating environment of A Schulman's laboratory (University of Helsinki), the initial discovery was made. A Mikic from IFVC, Novi Sad, Serbia, and E Smrzova from Prague Botanic Garden, are acknowledged for kindly providing various legume specimens.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Bagshaw ATM, Pitt JPW, Gemmell NJ (2008). High frequency of microsatellites in *S. cerevisiae* meiotic recombination hotspots. *BMC Genomics* **9**: 49.

Buschiazzo E, Gemmell NJ (2006). The rise, fall and renaissance of microsatellites in eukaryotic genomes. *Bioessays* **28**: 1040–1050.

Flavell AJ, Bolshakov VN, Booth A, Jing R, Russell J, Ellis NTH *et al.* (2003). A microarray-based high throughput molecular marker genotyping method: the tagged microarray marker (TAM) approach. *Nucl Acid Res* **31**: e115.

Gao L, Xu H (2008). Comparisons of mutation rate variation at genome-wide microsatellites: evolutionary insights from two cultivated rice and their wild relatives. *BMC Evol Biol* **8**: 1–14.

Gutierrez MV, Vaz Patto MC, Huguet T, Cubero JI, Moreno MT, Torres AM (2005). Cross-species amplification of *Medicago truncatula* microsatellites across three major pulse crops. *Theor Appl Genet* **110**: 1210–1217.

Han ZJ, de Lanerolle P (2008). Naturally extended CT AG repeats increase H-DNA structures and promoter activity in the smooth muscle myosin light chain kinase gene. *Mol Cell Biol* **28**: 863–872.

Inukai T (2004). Role of transposable elements in the propagation of minisatellites in the rice genome. *Mol Genet Genomics* **271**: 220–227.

Jing R, Knox MR, Lee JM, Vershinin AV, Ambrose M, Ellis TH *et al.* (2005). Insertional polymorphism and antiquity of PDR1 retrotransposon insertions in *Pisum* species. *Genetics* **171**: 741–752.

Johnson PC, Webster LM, Adam A, Buckland R, Dawson DA, Keller LF (2006). Abundant variation in microsatellites of the parasitic nematode *Trichostrongylus tenuis* and linkage to a tandem repeat. *Mol Biochem Parasitol* **148**: 210–218.

Kalendar R, Schulman AH (2006). IRAP and REMAP for retrotransposon-based genotyping and fingerprinting. *Nat Protoc* **1**: 2478–2484.

Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK *et al.* (2007). Human genome ultraconserved elements are ultraselected. *Science* **317**: 915–916.

Kazazian Jr HH (2004). Mobile elements: drivers of genome evolution. *Science* **303**: 1626–1632.

Koike M, Kawaura K Ogihara Y, Torada A (2006). Isolation and characterization of SSR sequences from the genome and TAC clones of common wheat using the PCR technique. *Genome* **49**: 432–444.

Loridon K, McPhee K, Morin J, Dubreuil P, Pilet-Nayel ML, Aubert G *et al.* (2005). Microsatellite marker polymorphism and mapping in pea (*Pisum sativum* L.). *Theor Appl Genet* **111**: 1022–1031.

Lowe TM, Eddy SR (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucl Acid Res* **25**: 955–964.

Lu Q, Teare JM, Granok H, Swede MJ, Xu J, Elgin SCR (2003). The capacity to form H-DNA cannot substitute for GAGA factor binding to a (CT)n (GA)n regulatory site. *Nucl Acid Res* **31**: 2483–2494.

Macas J, Neumann P, Navrátilová A (2007). Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Genomics* **21**: 427.

Meglecz E, Anderson SJ, Bourguet D, Butcher R, Caldas A, Cassel-Lundhagen A *et al.* (2007). Microsatellite flanking region similarities among different loci within insect species. *Insect Mol Biol* **16**: 175–185.

Morgante M, Hanafey M, Powell W (2002). Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nature Genet* **30**: 194–200.

Mun JH, Kim DJ, Choi HK, Gish J, Debellé F, Mudge J *et al.* (2006). Distribution of microsatellites in the genome of *Medicago truncatula*: A resource of genetic markers that integrate genetic and physical maps. *Genetics* **172**: 2541–2555.

Neumann P, Pozarkova D, Macas J (2003). Highly abundant pea LTR-retrotransposon Ogre is constitutively transcribed and partially spliced. *Plant Mol Biol* **53**: 399–410.

Noma K, Nakajima R, Ohtsubo H, Ohtsubo E (1997). RIRE-1, a retrotransposon from wild rice *Oryza australiensis*. *Genes Genet Syst* **72**: 131–140.

Orlov YL, Potapov VN (2004). Complexity: an internet resource for analysis of DNA sequence complexity. *Nucl Acid Res* **32**: 628–633.

Pandian A, Ford R, Taylor PWJ (2000). Transferability of sequence tagged microsatellite site (STMS) primers across four major pulses. *Plant Mol Biol Rep* **18**: 1–8.

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H *et al.* (2006). Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. *Genome Res* **16**: 1262–1269.

Ramsay L, Macaulay M, Cardle L, Morgante M, degli Ivanissevich S, Maestri E *et al.* (1999). Intimate association of microsatellite repeats with retrotransposons and other dispersed repetitive elements in barley. *Plant J* **17**: 415–425.

Roulin A, Piegu B, Wing RA, Panaud O (2008). Evidence of multiple horizontal transfers of the long terminal repeat retrotransposon RIRE-1 within the genus *Oryza*. *Plant J* **53**: 950–959.

Rubinsztein DC (1999). Trinucleotide expansion mutations cause diseases which do not conform to classical Mendelian expectations. In: Goldstein DB, Schlötterer C (eds). *Microsatellites: evolution and applications*. Oxford University Press: Oxford, UK. pp 80–97.

Schlötterer C (2000). Evolutionary dynamics of microsatellite DNA. *Chromosoma* **109**: 365–371.

Schneeberger RG, Zhang K, Tatarinova T, Troukhan M, Kwok SF, Drais J *et al.* (2005). *Agrobacterium* T-DNA integration in *Arabidopsis* is correlated with DNA sequence composition that occurs frequently in gene promoter regions. *Funct Integr Genomics* **5**: 240–253.

Smýkal P (2006). Development of an efficient retrotransposon-based fingerprinting method for rapid pea variety identification. *J Appl Genetics* **47**: 221–230.

Smýkal P, Hýbl M, Corander J, Jarkovský J, Flavell AJ, Griga M (2008). Genetic diversity and population structure of pea (*Pisum sativum* L.) varieties derived from combined retrotransposon, microsatellite and morphological marker analysis. *Theor Appl Genet* **117**: 413–424.

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

Tautz D, Renz M (1984). Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucl Acid Res* **12**: 4127–4138.

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001). Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res* **11**: 1441–1452.

Tero N, Neumeier H, Gudavalli R, Schlötterer C (2006). *Silene tatarica* microsatellites are frequently located in repetitive DNA. *J Evol Biol* **19**: 1612–1619.

Van't Hof AE, Brakefield PM, Saccheri IJ, Zwaan BJ (2007). Evolutionary dynamics of multilocus microsatellite arrangements in the genome of the butterfly *Bicyclus anynana*, with implications for other *Lepidoptera*. *Heredity* **98**: 320–328.

Vicient CM, Kalendar R, Anamthawat-Jónsson K, Schulman AH (1999). Structure, functionality, and evolution of the BARE-1 retrotransposon of barley. *Genetica* **107**: 53–63.

Vicient CM, Kalendar R, Schulman AH (2005). Variability, recombination, and mosaic evolution of the barley BARE-1 retrotransposon. *J Mol Evol* **61**: 275–291.

Vitte C, Bennetzen JL (2006). Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. *Proc Natl Acad Sci USA* **103**: 17638–17643.

Von Stackelberg M, Rensing SA, Reski R (2008). Identification of genic moss SSR markers and a comparative analysis of twenty-four algal and plant gene indices reveal species-specific rather than group-specific characteristics of microsatellites. *BMC Plant Biol* **6**: 9.

Wicker T, Keller B (2007). Genome-wide comparative analysis of *copia* retrotransposons in *Triticeae*, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. *Genome Res* **17**: 1072–1081.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B *et al*. (2007). A unified classification system for eukaryotic transposable elements. *Nature Rev Genet* **8**: 973–982.

Wilder J, Hollocher H (2001). Mobile elements and the genesis of microsatellites in dipterans. *Mol Biol Evol* **18**: 384–392.

Zhang L, Zuo K, Zhang F, Cao Y, Wang J, Zhang Y *et al*. (2006). Conservation of noncoding microsatellites in plants: implication for gene regulation. *BMC Genomics* **7**: 323.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)