EVOLUTIONARY RELATIONSHIPS AND DISTRIBUTION OF NON-LTR RETROTRANSPOSONS IN EUKARYOTES

Beresikov E.*, Novikova O.*, Makarevich I., Lashina V., Plasterk R., Blinov A.

1 Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; 2 Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, USA; 3 Hubrecht Laboratory, Netherlands Institute for Developmental Biology, Utrecht, The Netherlands

* Corresponding author: e-mail: novikova@bionet.nsc.ru

Keywords: mobile elements, non-LTR retrotransposons, distribution, evolution, computer analysis, tree of life

Resume

Non-LTR retrotransposons constitute a substantial part of some eukaryotic genomes and can play an important role in such evolutionary processes as genome rearrangement, gene formation and transcriptional patterning. Phylogenetic relationships between different families of non-LTR retrotransposons, based on sequences of reverse transcriptase, endonuclease and RNase H domains, are well-established, but no systematic study on the distribution of non-LTR elements among living organisms has been performed to date. We have screened 30 species, which represent all major eukaryotic phyla, for the presence of 19 different families of non-LTR retrotransposons in their genomes using PCR with family-specific consensus-degenerate hybrid oligonucleotide primers. In addition, we have used our previously developed computational approach to search for non-LTR reverse transcriptase sequences in the data generated by 53 different genome sequencing projects. The combination of the two approaches resulted in the identification of 87 new non-LTR elements, belonging mostly to known families, in 44 species, thus doubling the number of species with characterized non-LTR retrotransposon content. The data obtained in this study allowed us to project the known phylogeny of non-LTR elements onto the Tree of Life, coupling the evolution of non-LTR retrotransposons with the evolution of their host genomes. Our findings support the principle steps of the previously proposed scenario for the evolution of non-LTR elements, such as order of acquisition of enzymatic domains and vertical mode of inheritance, but also provide additional resolution power to the current understanding of the evolution of non-LTR elements. Based on the structural, phylogenetic and distribution data, we propose a refined taxonomy of non-LTR retrotransposable elements.

Introduction

Retrotransposons are mobile genetic elements that propagate themselves by reverse transcription of an RNA intermediate. There are two major classes of retrotransposons, which differ structurally and mechanistically: LTR retrotransposons possess long terminal repeats (LTRs) and have a transposition mechanism similar to that of retroviruses, whereas non-long terminal repeat (non-LTR) retrotransposable elements do not have terminal repeats and utilize a simpler target-primed reverse transcription (TPRT) mechanism for their retrotransposition. In TPRT, the element-encoded endonuclease cleaves the genomic DNA, the reverse transcriptase uses this break to prime reverse transcription from the element RNA, and the resulting cDNA copy is then integrated into the target site (Luan et al., 1993). SINE (short interspersed nucleotide elements) elements use the non-LTR retrotransposon machinery for their transposition (Ogiwara et al., 2002). Non-LTR retrotransposons have been found in all eukaryotes investigated to date and are the most abundant class of transposable elements. The copy number of the elements may vary from several copies per genome, as has been shown for some elements in Drosophila melanogaster (Berezikov et al., 2000), to more than 800,000 copies (~20% of the genome) for L1
elements in human (International Human Genome Sequencing Consortium, 2001). Transposition of non-LTR elements causes a hybrid disgenesis in *Drosophila* (Fawcett et al., 1986) and genetic diseases in human (Kazazian, Moran, 1998), and has been implicated in the emergence of pseudogenes (Esnault et al., 2000) and exon shuffling (Moran et al., 1999). Thus, non-LTR retrotransposons play important roles in the structural organization and evolution of the genomes they inhabit.

Phylogenetic analysis of non-LTR retrotransposons based on the reverse transcriptases domains allowed to distinguish 15 phylogenetic clades. Based on structural and phylogenetic features of different elements, Malik, Burke and Eickbush (1999) developed a scenario for evolution of non-LTR retrotransposons and demonstrated that non-LTR elements are inherited strictly by vertical transmission. Only a few cases of possible horizontal transfer of non-LTR retrotransposons have been suggested in literature (Kordis, Gubensek, 1999). According to the scenario of Malik, Burke and Eickbush, the most ancient clades of non-LTR retrotransposons (GENIE, CRE, R2, NeSL-1, and R4) contain only one ORF and show site-specific distribution in the genomes (Malik et al., 1999; Malik, Eickbush, 2000), which is provided by restriction-enzyme-like endonucleases (REL-end) these elements encode. During further evolution of mobile elements, the REL-end domain was substituted with an apurinic/apyrimidinic (AP) endonuclease acquired from the host cells. All younger clades (L1, RTE, Tad, R1, LOA, I, Jockey, CR1, Rex1, and L2) possess the AP endonuclease domain. The acquisition of the AP endonuclease resulted in losing target site specificity for all the elements (except the R1 clade and some elements from the L1 clade), and coincided with the origin of a second ORF in front of the RT-encoding ORF. Finally, elements of some clades obtained one more enzymatic domain in the second ORF – the RNase H domain.

It is clear that the presented distribution of the clades among living organisms reflects the extent of non-LTR retrotransposons investigation. Thus, representatives of nine out of fifteen known clades have been found in insects (mainly in dipteran species), where mobile elements were extensively characterized. A new approach for a broad investigation of the non-LTR retrotransposons distribution and evolution relies on a search for new transposable elements using the degenerate primers specific for the most conservative parts of the elements (Lovsin et al., 2001). Using this approach, detailed distribution of non-LTR retrotransposons has been analyzed for main taxa of the phylum Metazoa (Archipova, Meselson, 2000). Despite a large number of non-LTR retritransposon studies, no complete characterization of the distribution of these elements among different phyla in Eukaryotes has been made to date. Such characterization would provide new insights into the origin of non-LTR elements, their evolution and dispersal among living organisms, and also into the functional roles that non-LTR retrotransposons may play in multiple evolutionary processes, such as speciation, genome rearrangements, etc.

**Materials and Methods**

**Total DNA isolation.** Total DNA was isolated as described previously (Guryev et al., 2001).

**PCR amplification and sequencing.** Based on the comparison of 72 sequences of non-LTR retrotransposon reverse transcriptase domains, degenerate oligonucleotides primers were constructed using CODEHOPE software (Rose et al., 1998) to amplify a 500 bp region of reverse transcriptase. In total, eleven sense and eleven anti-sense degenerate primers were selected. Nineteen combinations of these primers were unique to nineteen selected families of non-LTR retrotransposons PCR amplification was performed using 0.1 µg of genomic DNA in 10-µl volume of 10 mM Tris-HCl (pH8.9), 1 mM (NH4)2SO4, 1.5 mM MgCl2, 200 µM each of four dNTPs, 0.5 µM primers, and 2.5 units of Taq polymerase. After an initial denaturation step for 3 min at 94 °C, the PCR reactions were subjected to 30 cycles of amplification consisting of 30 sec denaturation at 94°C, 42 sec annealing at 52 °C, and 1 min extension at 72 °C. PCR results were assayed by agarose gel electrophoresis and PCR fragments of expected size were cloned into a pBlueScript (KS+) vector using standard

---

182
procedures. The inserts were sequenced using DyeNamic ET chemistry on an ABI 3700 sequencer.

**Sequence analysis.** Search for non-LTR elements in publicly available genome sequencing data was performed as described previously (Berezikov et al., 2000). The newly identified RT sequences were aligned to the previously established alignment of non-LTR reverse transcriptases (Malik et al., 1999) using Clustal W software. The alignment used for phylogenetic tree construction is available as supplementary material. Phylogenetic trees were generated by Neighbor-Joining method using MEGA2 software package (Kumar et al., 2001).

**Results and Discussion**

**Screening of representative species by PCR with degenerate primers.** For amplification of RT sequences from different organisms, we designed degenerate primers using the consensus-degenerate hybrid oligonucleotide primers approach (Rose et al., 1998). These primers have a short degenerate 3' core region and a longer consensus 5' region, allowing amplification of distantly related sequences. Only 3 to 4 conserved amino acid residues in the analyzed group of sequences are essential for primer design. We have analyzed alignment of the RT sequences from 72 non-LTR elements (Malik et al., 1999) and distinguished 19 families of retrotransposons based on the identity of amino acid motifs in the most conserved regions of the alignment. To investigate distribution of the families of non-LTR retrotransposons among living organisms, we selected thirty representative species that cover main eukaryotic taxa. The nineteen primer combinations were used to screen genomes of the thirty selected species for the presence of different families of non-LTR retrotransposons. The results of PCR were considered positive if a band of expected size (~500 bp) was observed. The PCR screening results were in good agreement with the known data of non-LTR retrotransposons distribution. To determine the nature of PCR fragments amplified by degenerate primers, we cloned and sequenced PCR products from one or several species for most of the non-LTR retrotransposon families. In total, 197 clones belonging to the different families were sequenced. Sequencing of the clones confirmed that degenerate PCR primers amplified the actual RT domains of non-LTR retrotransposons.

**Computational screen for non-LTR reverse transcriptase sequences.** Data generated by genome sequencing projects can provide definitive information on the structure and distribution of transposable elements among living organisms. We used our previously developed computational approach (Berezikov et al., 2000) to screen data, generated by various sequencing projects, for non-LTR retrotransposons. The approach uses the profile hidden Markov model (HMM) software to find sequences matching the reverse transcriptase model and containing the motif $F(Y)XDD$, which is conserved among all reverse transcriptases. At this step, most of the potential reverse transcriptase sequences are identified. Next, BLAST analysis is performed to group the sequences by homology to reverse transcriptases of LTR or non-LTR elements, telomerases or retroviruses. Finally, redundancy of sequences in each group is removed and phylogenetic tree is constructed to estimate relationships between newly identified and known reverse transcriptases. We analyzed sequences of 53 eukaryotic organisms. Altogether, using our computational approach we identified 70 different non-LTR reverse transcriptase sequences in 35 species. Among these sequences, 29 directly corresponded to already known elements, whereas 41 sequences were new reverse transcriptases from 22 previously uncharacterized species. Most of the newly identified RTs clearly grouped with the known elements on a phylogenetic tree and covered all the clades of non-LTR retrotransposons.

**Distribution of different non-LTR retrotransposon families among living organisms and their evolutionary relationships.** The evolutionary scenario for non-LTR retrotransposons proposed by Malik et al. (1999) was based on a large but insect-biased dataset of non-LTR elements. Moreover, additional clades of retrotransposons have been described since the publication of their work. Our systematic PCR screening for non-LTR elements in representative species from major phylogenetic groups, as well as computational analysis of sequenced genomes, have substantially increased the
number of species and elements in a dataset, thus allowing us to propose a refined model of the evolution and distribution of different families of non-LTR retrotransposons among eukaryotes. It should be noted that our model is principally the same as the model developed by Malik et al. (1999), and strongly relies on their finding that non-LTR elements evolve mainly through vertical transmission, and not through horizontal transfer. However, besides the phylogeny of the retrotransposon sequences themselves, our model also incorporates the phylogeny of host species, thus providing additional resolution power and a more integrative view for the evolution of non-LTR elements. Finally, we would like to present our version of the taxonomy of non-LTR retrotransposable elements that is based on all available data on the structural organization and evolutionary relationships of these elements. All non-LTR elements are divided into three groups according to the presence of one or two ORFs and AP and RNAseH domains. The first group of elements containing only one ORF includes 5 clades (GENIE, CRE, R4, NeSL-1, and R2) with one family in each clade. The second group of non-LTR retrotransposons containing two ORFs and the AP endonuclease domain also includes 5 clades (L1, RTE, CR1, Rex1, and Jockey) that together comprise 15 families. The last group containing two ORFs, AP endonuclease and RNaseH domains includes two clades, Tad and I, with three families in each of these clades. The taxonomy proposed here is most likely not the final version and will be changing with discoveries of new elements. However, the major division of non-LTR elements into clades will probably remain unchanged.

References


