Published in final edited form as:

Gene. 2009 December 15; 448(2): 207–213. doi:10.1016/j.gene.2009.07.019.

Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences

Vladimir V. Kapitonov^{*}, Sébastien Tempel, and Jerzy Jurka^{*}

Genetic Information Research Institute, Mountain View, California

Abstract

Rapidly growing number of sequenced genomes requires fast and accurate computational tools for analysis of different transposable elements (TEs). In this paper we focus on rapid and reliable procedure for classification of autonomous non-LTR retrotransposons based on alignment and clustering of their reverse transcriptase (RT) domains. Typically, the RT domain protein sequences encoded by different non-LTR retrotransposons are similar to each other in terms of significant BLASTP E-values. Therefore, they can be easily detected by the routine BLASTP searches of genomic DNA sequences coding for proteins similar to the RT domains of known non-LTR retrotransposons. However, detailed classification of non-LTR retrotransposons, i.e. their assignment to specific clades, is a slow and complex procedure that is not formalized or integrated as a standard set of computational methods and data. Here we describe a tool (RTclass1) designed for the fast and accurate automated assignment of novel non-LTR retrotransposons to known or novel clades using phylogenetic analysis of the RT domain protein sequences. RTclass1 classifies a particular non-LTR retrotransposon based on its RT domain in less than 10 minutes on a standard desktop computer and achieves 99.5% accuracy. RTclass1 works either as a standalone program installed locally or as a web-server that can be accessed distantly by uploading sequence data through the internet (<http://www.girinst.org/RTphylogeny/RTclass1>).

Keywords

transposable elements; non-LTR retrotransposons; classification; phylogenetic analysis; genome annotation

1. Introduction

All eukaryotic transposable elements (TEs) belong to only two types: retrotransposons and DNA transposons (Craig et al., 2002; Jurka et al., 2007; Kapitonov and Jurka, 2008). All genomic and extra-chromosomal copies of retrotransposons are transposed through an RNA intermediate. Their messenger RNA (mRNA) is expressed in the host cell, reverse transcribed, and the resulting DNA copy (cDNA) is integrated into the host genome. Reverse transcription and integration steps are catalyzed by reverse transcriptase (RT) and endonuclease/integrase (EN/IN), which are encoded by autonomous retrotransposons. Unlike retrotransposons, DNA

© 2009 Elsevier B.V. All rights reserved.

^{*}corresponding authors: Vladimir V. Kapitonov, Genetic Information Research Institute, 1925 Landings Dr, CA 94041, USA, vladimir@girinst.org, Jerzy Jurka, Genetic Information Research Institute, 1925 Landings Dr, CA 94041, USA, jurka@girinst.org.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

transposons are transposed by transferring their copies from one chromosomal location to another without copying their RNA intermediates. DNA transpositions are catalyzed by DNA transposases encoded by autonomous DNA transposons (Craig et al., 2002).

Eukaryotic retrotransposons can be divided into four classes: non-long terminal repeat (non-LTR) retrotransposons, LTR retrotransposons, *Penelope*, and *DIRS* retrotransposons. While the first two classes are well established and studied (Eickbush and Malik, 2002), the *Penelope* and *DIRS* classes were only recently introduced (Arkhipova et al., 2003; Evgen'ev and Arkhipova, 2005; Poulter and Goodwin, 2005; Lorenzi et al., 2006; Gladyshev and Arkhipova, 2007). Members of all the four retrotransposon classes are present in the genomes of all eukaryotic kingdoms: Protista, Plantae, Fungi, and Animalia.

A typical autonomous non-LTR retrotransposon generally referred to as LINE (Long INterspersed Element) contains one or two open reading frames (ORFs), and an internal RNA polymeraseII promoter in its 5'-terminal region that drives transcription of the full-length retrotransposon. Both its RT and EN domains are universally encoded by the same ORF (Eickbush and Malik, 2002). An mRNA expressed during transcription of a genomic non-LTR retrotransposon serves as a template for reverse transcription, and the resulting cDNA is inserted in the genome. The mechanism of retrotransposition and integration of LINEs into the genome is viewed as a process called target-primed reverse transcription (TPRT) (Luan et al., 1993; Eickbush and Malik, 2002). According to the TPRT model, reverse transcription is primed by the free 3' hydroxyl group at the target DNA nick introduced by EN.

Despite the basic common mechanism, there are some variations in target preferences, duplications/deletions at the insertion sites as well as in the speed and accuracy of reverse transcription. Such variations often correlate with sequence differences among proteins encoded by different phylogenetic groups of non-LTR retrotransposons (Eickbush and Malik, 2002). Therefore, meaningful classification is an important step in studies of LINE elements.

Here we describe a simple approach to produce a semi-automatic classification of autonomous non-LTR retrotransposons based on phylogenetic analysis of their RT domain protein sequences. In modern taxonomy, a monophyletic group of living or fossil organisms that consists of a single common ancestor and all its descendants is often referred to as a clade (from *klados* or "branch" in ancient Greek). The term clade was introduced in 1959 by Julian Huxley (Huxley, 1959), and became popular in evolutionary biology during the last 20 years. In 1999, Malik, Burke and Eickbush proposed the use of the term "clade" to represent those non-LTR retrotransposons that share the same structural features, are grouped together based on phylogenetic analysis of the reverse transcriptase domain, and dated back to the Precambrian era (Malik et al., 1999).

Based on structural features of non-LTR retrotransposons and phylogeny of RTs, they were assigned to five groups, called *R2*, *L1*, *RTE*, *I*, and *Jockey*, which were subdivided prior 2003 into 15 clades: CRE, NeSL, R4, R2, L1, RTE, Tad1, R2, LOA, I, Ingi, Jockey, CR1, Rex1, and L2 (Malik et al., 1999; Lovsin et al., 2001; Eickbush and Malik, 2002). Another nine clades, including Outcast, Hero, RandI (also known as Dualen), Daphne, Tx1, RTE, Proto1 and Proto2 have been reported since 2003 (see table 1). Currently, we consider 28 different clades, including the L2A, L2B, Nimb and RTETP clades introduced in this manuscript (figure 1, table 1, and supplemental figure 2). It is believed that the *R2* group is composed of the most ancient non-LTR retrotransposons: the *CRE*, *NeSL*, *R2*, *Hero*, and *R4* clades, which are characterized by a single ORF coding for the RT and EN domains. The endonuclease domain is similar to different restriction enzymes and is always preceded by the RT domain. Most likely, the restriction-like endonuclease (RLE) in retrotransposons from the *R2* group is responsible for their frequent target-site specificity (Kojima and Fujiwara, 2005b). Members of the *L1*, *RTE*,

I, Jockey groups encode the apurinic-apyrimidinic endonuclease (APE), which is always N terminal to the RT domain. In addition to RT and EN, all retrotransposons from the I group code for ribonuclease (RNase) H (Eickbush and Malik, 2002). Analogously, diverse plant L1 retrotransposons also code for the RNase H (V.K., unpublished data). Usually, Non-LTR retrotransposons are transmitted vertically, with only few exceptions in the RTE clade (Kordis and Gubensek, 1997). *RandI*/Dualen retrotransposons identified in the green algae form an ancient clade; like R2-group elements, they contain only one ORF that codes for a protein with the RLE domain (although its similarity to known RLEs is marginal) (Kojima and Fujiwara, 2005a). In addition, this protein contains the conserved APE domain (Kapitonov and Jurka, 2004; Kojima and Fujiwara, 2005a). Therefore, we consider the *RandI* clade as a founder of a new group of non-LTR retrotransposons (figure 1 and supplemental figure 2).

The RT domain is functionally the most important and the only domain present universally in all autonomous non-LTR retrotransposons (Eickbush and Malik, 2002) (see also figure 1). Moreover, numerous studies by independent groups devoted to the assignment of non-LTR retrotransposons to different clades have relied basically on phylogeny of their RT domains and produced results that seem to be quite stable and reliable despite the amount of new data accumulated after publications (Malik et al., 1999; Lovsin et al., 2001; Eickbush and Malik, 2002; Kojima and Fujiwara, 2004; Putnam et al., 2007). Therefore, the RT-based phylogeny is probably unavoidable and mostly sufficient approach for assignment of diverse retrotransposons to known clades of non-LTR retrotransposons. However, the current methods of robust phylogenetic analysis are extremely slow. A construction of a reliable tree for some 100 protein sequences often takes more than a day, especially if these sequences are less than 20% identical to each other, as is typical for the RT domain of non-LTR retrotransposons (Fig. 2). Moreover, selection of diverse protein sequences encoded by different families of non-LTR retrotransposons is crucial for obtaining reliable results regarding classification of novel retrotransposons. Another problem is a huge diversity and complexity of modern methods of phylogenetic analysis (<http://evolution.genetics.washington.edu/phylip/software.html>). As a result, the classification of novel retrotransposons can be either inaccurate or unreasonably time consuming. Therefore, the use of well established reference sets of protein sequences encoded by previously classified non-LTR retrotransposons and of reliable and fast methods of phylogenetic analysis are highly important as “the rosetta stone” in future studies induced by an explosion of sequence data.

2. Materials and methods

A basic scheme depicting an assignment of a protein sequence to a specific clade of non-LTR retrotransposons is outlined in Figure 3. Hereafter, we use the term “classification” as a synonym of the assignment to a known or new clade. First, a set of protein sequences of the RT domain encoded by known classified non-LTR retrotransposons was collected from Repbase (Jurka et al., 2005) and GenBank. Our choice of sequences included in this collection, named the “RTclass1 learning set”, was limited by two self-imposed restrictions: (i) the protein sequence identity between any two sequences included in the collection must be $\leq 60\%$, and (ii) the collection must contain only currently active or young non-LTR retrotransposons, represented mostly by their consensus sequences. The first restriction forces us to increase the amount of useful information in the learning set not just by the increase of the number of sequences but rather by the increase of the RT protein diversity covered by the included sequences. Most likely, inclusion of numerous sequences highly identical to each other would lead to dramatically slow computations, without significant improvement of the classification accuracy. The second restriction minimizes the amount of background noise introduced in the learning set due to numerous “dead mutations” accumulated in genomic copies of non-LTR retrotransposons that lost their mobility many million years ago. The average pairwise protein sequence identity between RTs that belong to two different clades is only 19% (Fig. 2).

Therefore, even a small number of “dead mutations” included in the learning set may lead to errors in the multiple alignment and wrong classification. We consider a particular family of non-LTR retrotransposon to be young as long as the host genome contains several members/copies of this family that are less than 10% divergent from each other. The consensus sequence built from a multiple alignment of the genomic copies should be free of numerous “dead mutations”. Sometimes, the genome contains only a single copy of a particular family of non-LTR retrotransposons. We consider this single copy retrotransposon young if it codes for the standard-size ORFs without stop-codons interrupting them. Given that the standard ORF encoding the RT-containing protein is longer than 3 kb, the absence of stop-codons in such a long sequence ensures small number of “dead mutations” accumulated in the retrotransposon. As a result, the current RTclass1 learning set consists of 211 protein sequences of the RT domain from diverse families of classified non-LTR retrotransposons that belong to all known clades (see supplemental figures S1 and figure S2).

When a protein sequence encoded by a new retrotransposon is taken for classification (Fig. 3), boundaries of its RT domain are defined based on BLASTP similarities to the RT domain sequences collected in the learning set. In the next step, the multiple alignment of the new RT sequence with all sequences from the learning set is obtained by using MUSCLE (Edgar, 2004a;Edgar, 2004b). Given that the learning set was composed of $N=211$ sequences, we have tested whether the multiple alignment of $N+1$ sequences could be replaced by the realignment of the new unclassified sequence with the existing multiple alignment of the N sequences. In such realignment, the multiple alignment of N sequences can be only modified by indels introduced simultaneously at the same positions in all N sequences. Such an addition of a new sequence to the old multiple alignment, also known as the “profile alignment,” is implemented in CLUSTAL (Larkin et al., 2007) as well as MUSCLE (Edgar, 2004a) and is extremely fast. Unfortunately, the accuracy of the profile alignment of highly diverse RT domain sequences encoded by non-LTR retrotransposons is not adequate. For instance, we prepared a random sample of 15 non-LTR retrotransposons that were not included in the learning set. In seven (out of the 15) retrotransposons, the “profile alignment”-based classification differed from the expected classification supported by the standard multiple alignment. Therefore, we cannot rely on the profile alignment.

Nevertheless, according to the previously reported estimates, execution time of standard multiple alignment by MUSCLE increases only tenfold as the number of ~ 300 -aa protein sequences increases from 200 to 1000 (Edgar, 2004a). Currently, the multiple alignment of the 211 RT domain sequences from the learning set takes ~ 10 seconds. Therefore, a multiple alignment of an expanded set of 1000 RT sequences (this is the expected number of sequences included in the learning set in the next two years) will take less than 2 minutes.

3. Results

3.1. Choosing the method of phylogenetic analysis

Our main objective is to develop a fast and reliable method that would permit to assign unclassified non-LTR retrotransposons to known and novel clades, either locally through a pipeline installed on a standard desktop computer or distantly via a web-server. Here, we present a simple procedure for ranking different methods and programs developed specifically for fast phylogenetic analysis of thousands of proteins sequences, including BIONJ (Gascuel, 2000), Clearcut (Sheneman et al., 2006), FastME (Desper and Gascuel, 2002), PHYLO_WIN (Galtier et al., 1996), QuickTree (Howe et al., 2002) and RaxML (Stamatakis et al., 2005).

For each method listed above, we used the same multiple alignment of previously classified 100 RT domains representing established clades of non-LTR retrotransposons, which we collected during recent identification and classification of non-LTR retrotransposons in the

Nematostella vectensis genome (Putnam et al., 2007). The model phylogenetic tree of RT domain sequences encoded by these retrotransposons was constructed by using MEGA4 (Tamura et al., 2007). This tree was also supported by numerous studies of non-LTR retrotransposons in the past (Eickbush and Malik, 2002; Kojima and Fujiwara, 2004; Kojima and Fujiwara, 2005a). In the model tree, which represented 17 different clades of non-LTR retrotransposons, all sequence names were grouped into 17 model clusters, where each cluster was composed of the names of sequences that belonged to the same clade.

For each tested method we created 1000 bootstrap trees by generating permutations in the original multiple alignment by SEQBOOT from the PHYLIP package (Felsenstein, 2005). Every bootstrap tree, based on its Newick format representation (<http://evolution.genetics.washington.edu/phylip/newicktree.html>), was automatically split into all possible clusters. A cluster was defined as a Newick substring bordered by the left and right parentheses at its left and right ends and containing equal numbers of left and right parentheses, including the two border parentheses. As a result, every bootstrap cluster contained unique sequence names, and all clusters together contained the complete set of 100 sequence names. In a set of all possible clusters identified in the bootstrap tree, we kept only those 17 clusters that were closest to the 17 model tree clusters.

To determine how close was each cluster in the bootstrap tree to a particular model cluster, we counted the following numbers: the number of sequence names from the model cluster that were not present in the bootstrap cluster ($n1$), and the number of sequence names in the bootstrap cluster that were not present in the corresponding model cluster ($n2$). Based on these two numbers, we calculated the error α of the consensus cluster as

$$\alpha = \frac{n1}{M} + \frac{n2}{T},$$

where M and T were the numbers of all sequence names constituting the model and consensus clusters, respectively. The smaller the α error, the closer is the bootstrap cluster to the model cluster. By screening all possible consensus clusters for every model cluster, we identified the unique consensus cluster characterized by the smallest error α .

For each bootstrap tree generated by the same tested method, after choosing all 17 best bootstrap tree clusters closest to the corresponding 17 model clusters/clades, we calculated the error δ of the tested method as

$$\delta = \sum_{i=1}^{17} \left(\frac{n1_i}{M_i} + \frac{n2_i}{T_i} \right),$$

where M_i and T_i were the numbers of sequence names present in the corresponding to each other model and predicted bootstrap cluster i , respectively. The error of the method was calculated as the mean value of δ in the 1000 bootstrap trees. Among all tested methods (Table 2), BIONJ had the lowest error and was chosen as the best method for assignment of novel non-LTR retrotransposons to known clades.

3.2. The RTclass1 dataset

The RTclass1 dataset is composed of 211 RT domain protein sequences that belong to 28 clades: CRE, R2, R4, RandI/Dualen, NeSL, Hero, Tx1, L1, Proto1, Proto2, RTETP, RTE, RTE, I, Nimb, Ingi, Outcast, Jockey, Crack, Daphne, L2, CR1, Rex1, L2A, L2B, Tad1, R1, and Loa (Table 1). All 28 clades are currently introduced into the classification scheme

implemented in Repbase (Jurka et al., 2005; Kapitonov and Jurka, 2008). To keep the nomenclature of individual non-LTR retrotransposons simple, we recommend the following standard rule for naming every novel non-LTR retrotransposon: “name of the clade”-“family number”_“species abbreviation” (Kapitonov and Jurka, 2008).

3.3. Basic scheme of the RTclass1 tool

The basic classification scheme implemented in RTclass1 is flexible and allows simple modifications by choosing different methods of multiple alignment, estimation of the protein distances, and inferring phylogenetic trees (Fig. 3). The input protein sequence can be assigned to one of the known or novel clades in less than 10 minutes either by submitting it to the RTclass1 web-server or by executing the stand-alone program locally on a standard desktop computer with Linux operating system. The classification output consists of several reports: (1) the name of the clade the input sequence belongs to, or if it cannot be classified the word “out-group” is displayed, indicating an unknown, potentially novel clade; (2) the consensus phylogenetic tree built from 1000 bootstrap trees that can be viewed by any standard web-browser; (3) the “real” tree built from the multiple alignment of the input RT sequence and the RTclass1 sequences; (4) the multiple alignment.

3.4. Classification algorithm of RTclass1

The analyzed protein sequence should be in the FASTA format. In the first step, the RTclass1 tool uses WU-BLAST/CENSOR (Kohany et al., 2006) to extract the RT domain from the analyzed sequence (Fig. 3). In the second step, RTclass1 creates the multiple alignment of the analyzed domain sequence and the RTclass1 dataset of RT domains. This multiple alignment is transformed by random bootstrap permutations via SEQBOOT (Felsenstein, 2005) into 1000 bootstrap multiple alignments, which are used later for calculation of 1000 protein distance matrixes by CLEARCUT (Sheneman et al., 2006). For each protein distance matrix, RTclass1 infers 1000 phylogeny trees by using BIONJ (Gascuel, 2000). In the next step, by analyzing the obtained 1000 bootstrap trees via Consense (Felsenstein, 2005), RTclass1 creates the consensus bootstrap tree and identifies the model cluster that contains the input sequence.

4. Discussion

4.1. Recommendation for the large-scale genome classification

A particular eukaryotic genome may contain non-LTR retrotransposons that belong to more than 100 families (Putnam et al., 2007; Putnam et al., 2008). The phylogeny-based classification of all these families, even in its simplest version described here, takes more than 16 hours on an average desktop computer and demands hours of manual work. On the other hand, the phylogenetic analysis appears not to be necessary for classification of a non-LTR retrotransposon with RT domain over 40% identical to the domain encoded by some classified retrotransposon present in the learning set. This “magic” 40% threshold is well supported by the distribution of the inter-clade pairwise protein identity obtained for 211 RT sequences constituting the learning set (Fig. 2). In this set, the number of different pairs of sequences that belong to different clades equals 13780. The distribution of the pairwise protein identity of all these 13780 pairs is characterized by the mean and standard deviation equal 19.3% and 4.2%, respectively; with minimum and maximum values equal to 8% and 39%. Therefore, as long as the protein identity between two RT domain sequences is $\geq 40\%$, covering over 90% of their sequence length, one can safely assume that both sequences belong to the same clade.

The efficiency of this approach on a genome scale level can be demonstrated by our recent studies of non-LTR retrotransposons in the lancelet genome, which contains 192 families of non-LTR retrotransposons identified computationally (Putnam et al., 2008). As illustrated in figure 4, only 40 families need to be passed through the phylogenetic analysis described above

to obtain accurate classification. Out of all 192 families, 105 can be immediately assigned to known clades based on $\geq 40\%$ identity of their RT domain sequences to one of the classified RTs from the RTclass1 learning set (Fig. 4, step 1). Due to their dominant vertical transmission mode, most families of non-LTR retrotransposons present in a particular genome are much closer to each other rather than to non-LTR retrotransposons identified in other species, including those that constitute the RTclass1 learning set. Therefore, some of the 105 families classified based on high identities between their RTs to the RTclass1 sequences can be $\geq 40\%$ identical to the remaining 87 unclassified families. In fact, using these 105 sequences as a new learning set 2, we found that another 42 families can be classified based on high identities of their RT sequences to classified sequences from the set 2 (Fig. 4, step 2). Repeating iteratively the described procedure (Fig. 4, steps 3–4), we have only 40 families left that cannot be classified based on BLASTP identity of their RTs to the previously classified RTs.

4.2. Pitfalls of the RTclass1 tool

While the assignment of novel non-LTR retrotransposons by the RTclass1 tool is reliable and accurate, we would urge potential users of this tool to be cautious in inferring the macro-topology of the global tree of non-LTR retrotransposons, e.g. the topology of branches connecting different clades to each other. Given the low identity between RTs from different clades (Fig. 2), an accurate reconstruction of evolution of clades, especially the oldest ones, needs additional studies and sophisticated methods of phylogenetic analysis that would take into account both variations of the mutation rate at different amino acid positions of the RT domain and variations of the mutation rate in different species.

4.3. Future improvements

To improve the current classification procedure, we are planning to enhance it by analysis of other protein domains in non-LTR retrotransposons, including endonucleases, ribonuclease and ORF1-encoded proteins. In one year, we are going to increase significantly the number of diverse RT domain sequences from young non-LTR retrotransposons (from 211 to ~ 1000) constituting the RTclass1 dataset. We would also like to encourage a feedback from potential users, including requests for submissions of new sequences and clades in the RTclass1 dataset and Repbase.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Oleksiy Kohany for help with putting the RTclass1 tool to the web-server and Irina Arkhipova for valuable comments on the manuscript. This work was supported by the National Institutes of Health grant 5 P41 LM006252.

References

- Arkhipova IR, Pyatkov KI, Meselson M, Evgen'ev MB. Retroelements containing introns in diverse invertebrate taxa. *Nat Genet* 2003;33:123–124. [PubMed: 12524543]
- Biedler J, Tu Z. Non-LTR retrotransposons in the African malaria mosquito, *Anopheles gambiae*: unprecedented diversity and evidence of recent activity. *Mol Biol Evol* 2003;20:1811–1825. [PubMed: 12832632]
- Craig, NL.; Craigie, R.; Gellert, M.; Lambowitz, AM. *Mobile DNA II*. Washington, DC: ASM Press; 2002.
- Desper R, Gascuel O. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J Comput Biol* 2002;9:687–705. [PubMed: 12487758]

- Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004a;5:113. [PubMed: 15318951]
- Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004b;32:1792–1797. [PubMed: 15034147]
- Eickbush, TH.; Malik, HS. The age and evolution of non-LTR retrotransposable elements. In: Craig, NL.; Craigie, R.; Gellert, M.; Lambowitz, AM.; Craig, NL.; Craigie, R.; Gellert, M.; Lambowitz, AM.; Craig, NL.; Craigie, R.; Gellert, M.; Lambowitz, AMs, editors. *Mobile DNA II*. Washington DC: ASM Press; 2002. p. 1111–1144.
- Evgen'ev MB, Arkhipova IR. Penelope-like elements—a new class of retroelements: distribution, function and possible evolutionary significance. *Cytogenet Genome Res* 2005;110:510–521. [PubMed: 16093704]
- Felsenstein, J. PHYLIP ({Phylogeny Inference Package}) version 3.6. Distributed by the author. Seattle: Department of Genome Sciences, University of Washington; 2005.
- Galtier N, Gouy M, Gautier C. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* 1996;12:543–548. [PubMed: 9021275]
- Gascuel O. On the optimization principle in phylogenetic analysis and the minimum-evolution criterion. *Mol Biol Evol* 2000;17:401–405. [PubMed: 10723740]
- Gladyshev EA, Arkhipova IR. Telomere-associated endonuclease-deficient Penelope-like retroelements in diverse eukaryotes. *Proc Natl Acad Sci U S A* 2007;104:9352–9357. [PubMed: 17483479]
- Howe K, Bateman A, Durbin R. QuickTree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics* 2002;18:1546–1547. [PubMed: 12424131]
- Huxley, J. Function and taxonomic importance. London: Systematics Association; 1959. Clades and grades. Cain, A.J. (Cain, A.J.)Cain, A.J.s)
- Jurka J, Kapitonov VV, Kohany O, Jurka MV. Repetitive sequences in complex genomes: structure and evolution. *Annu Rev Genomics Hum Genet* 2007;8:241–259. [PubMed: 17506661]
- Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 2005;110:462–467. [PubMed: 16093699]
- Kapitonov VV, Jurka J. The esterase and PHD domains in CR1-like non-LTR retrotransposons. *Mol. Biol. Evol* 2003;20:38–46. [PubMed: 12519904]
- Kapitonov VV, Jurka J. RandI-1, a family of RandI non-LTR retrotransposons from the *Chlamydomonas reinhardtii* genome. *Repbases Reports* 2004;4:196.
- Kapitonov VV, Jurka J. CR1-12_SP, a family of non-LTR retrotransposons from the sea urchin genome. *Repbases Reports* 2005;5:70.
- Kapitonov VV, Jurka J. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat Rev Genet* 2008;9:411–412. author reply 414. [PubMed: 18421312]
- Kapitonov VV, Jurka J. Proto1 non-LTR retrotransposons from the *Naegleria gruberi* amoeboflagellate genome. *Repbases Reports* 2009a;9:1144–1148.
- Kapitonov VV, Jurka J. Proto2, a novel clade of metazoan non-LTR retrotransposons. *Repbases Reports* 2009b;9:1554–1563.
- Khazina E, Weichenrieder O. Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. *Proc Natl Acad Sci U S A* 2009;106:731–736. [PubMed: 19139409]
- Kohany O, Gentles AJ, Hankus L, Jurka J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* 2006;7:474. [PubMed: 17064419]
- Kojima KK, Fujiwara H. Cross-genome screening of novel sequence-specific non-LTR retrotransposons: various multicopy RNA genes and microsatellites are selected as targets. *Mol Biol Evol* 2004;21:207–217. [PubMed: 12949131]
- Kojima KK, Fujiwara H. An extraordinary retrotransposon family encoding dual endonucleases. *Genome Res* 2005a;15:1106–1117. [PubMed: 16077010]
- Kojima KK, Fujiwara H. Long-term inheritance of the 28S rDNA-specific retrotransposon R2. *Mol Biol Evol* 2005b;22:2157–2165. [PubMed: 16014872]
- Kojima KK, Kuma K, Toh H, Fujiwara H. Identification of rDNA-specific non-LTR retrotransposons in *Cnidaria*. *Mol Biol Evol* 2006;23:1984–1993. [PubMed: 16870681]

- Kordis D, Gubensek F. Bov-B long interspersed repeated DNA (LINE) sequences are present in *Vipera ammodytes* phospholipase A2 genes and in genomes of Viperidae snakes. *Eur J Biochem* 1997;246:772–779. [PubMed: 9219538]
- Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. Clustal W and Clustal X version 2.0. *Bioinformatics* 2007;23:2947–2948. [PubMed: 17846036]
- Lorenzi HA, Robledo G, Levin MJ. The VIPER elements of trypanosomes constitute a novel group of tyrosine recombinase-encoding retrotransposons. *Mol Biochem Parasitol* 2006;145:184–194. [PubMed: 16297462]
- Lovsin N, Gubensek F, Kordi D. Evolutionary dynamics in a novel L2 clade of non-LTR retrotransposons in Deuterostomia. *Mol Biol Evol* 2001;18:2213–2224. [PubMed: 11719571]
- Luan DD, Korman MH, Jakubczak JL, Eickbush TH. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* 1993;72:595–605. [PubMed: 7679954]
- Malik HS, Burke WD, Eickbush TH. The age and evolution of non-LTR retrotransposable elements. *Mol Biol Evol* 1999;16:793–805. [PubMed: 10368957]
- Malik HS, Eickbush TH. The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs. *Mol Biol Evol* 1998;15:1123–1134. [PubMed: 9729877]
- Poulter RT, Goodwin TJ. DIRS-1 and the other tyrosine recombinase retrotransposons. *Cytogenet Genome Res* 2005;110:575–588. [PubMed: 16093711]
- Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu JK, Benito-Gutierrez EL, Dubchak I, Garcia-Fernandez J, Gibson-Brown JJ, Grigoriev IV, Horton AC, de Jong PJ, Jurka J, Kapitonov VV, Kohara Y, Kuroki Y, Lindquist E, Lucas S, Osoegawa K, Pennacchio LA, Salamov AA, Satou Y, Sauka-Spengler T, Schmutz J, Shin IT, Toyoda A, Bronner-Fraser M, Fujiyama A, Holland LZ, Holland PW, Satoh N, Rokhsar DS. The amphioxus genome and the evolution of the chordate karyotype. *Nature* 2008;453:1064–1071. [PubMed: 18563158]
- Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV, Jurka J, Genikhovich G, Grigoriev IV, Lucas SM, Steele RE, Finnerty JR, Technau U, Martindale MQ, Rokhsar DS. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* 2007;317:86–94. [PubMed: 17615350]
- Schon I, Arkhipova IR. Two families of non-LTR retrotransposons, Syrinx and Daphne, from the Darwinulid ostracod, *Darwinula stevensoni*. *Gene* 2006;371:296–307. [PubMed: 16469453]
- Sheneman L, Evans J, Foster JA. Clearcut: a fast implementation of relaxed neighbor joining. *Bioinformatics* 2006;22:2823–2824. [PubMed: 16982706]
- Stamatakis A, Ludwig T, Meier H. RAXML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 2005;21:456–463. [PubMed: 15608047]
- Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 2007;24:1596–1599. [PubMed: 17488738]
- Volff JN, Korting C, Froschauer A, Sweeney K, Scharl M. Non-LTR retrotransposons encoding a restriction enzyme-like endonuclease in vertebrates. *J Mol Evol* 2001;52:351–360. [PubMed: 11343131]
- Volff JN, Korting C, Scharl M. Multiple lineages of the non-LTR retrotransposon Rex1 with varying success in invading fish genomes. *Mol Biol Evol* 2000;17:1673–1684. [PubMed: 11070055]

Fig. 1.

Schematic structure of non-LTR retrotransposons from different clades. The ORF1 and ORF2-encoded proteins are shown as short and long white rectangles. In the ORF2 proteins, black rectangles mark the RT domains; black and white asterisks stay for the APE and RLE endonucleases, respectively; scissors denote ribonuclease H. In the ORF1 proteins, bells and diamonds mark the esterase (Kapitonov and Jurka, 2003) and L1-like or RRM domains (Kapitonov and Jurka, 2005; Khazina and Weichenrieder, 2009). Also in ORF1 proteins, ovals indicate zinc knuckles: Cx2Cx4Hx4Cx5-8Cx2Cx3Hx4C. Domains and ORF1 that are present only in some families of a particular clade are in gray.

Fig. 2.

Histogram of pairwise protein identities (%) between any two RT domains from retrotransposons that belong to different clades. This histogram was obtained for the 211 RT sequences from classified non-LTR retrotransposons constituting the learning set.

Fig. 3.
Classification scheme implemented in RTclass1.

Fig. 4.
Iterative classification of non-LTR retrotransposons from the lancelet genome.

Table 1

Clades and non-LTR retrotransposons that constitute the RTclass1 dataset.

Clade	Non-LTR retrotransposons	Host species	References
CRE	CRE1, CRE2, CZAR, SLACS, Cn11, GilD, GilM, Cre-1_MB	protists	(Malik et al., 1999)
R2	R2_PS, PERERE-9, R2-1_SM, R2_DM, R2Ci-B, R2_AM, R2Dr, R2-1_PM, R2-2_PM, R2_LP, R2-1_TSP	protists, cnidarians, insects, planarian, sea squirt, fish, lamprey, nematode	(Malik et al., 1999)
RandI/ Dualen	RandI-1, RandI-4, RandI-6	green algae	(Kapitonov and Jurka, 2004; Kojima and Fujiwara, 2005a)
R4	EhRLE2, EhRLE3, DongAG, R4_AL, R4-1_ED, DONG_FR, R4-1_AC	protists, nematode, insects, fish, lizard	(Malik et al., 1999; Volff et al., 2001)
NeSL	NeSL-1, LIN9_SM, R2-1a_Cis, YURECi, R2I-2_PI, R2I-1_PI, R5-2_SM, R5-1_SM, R5, NeSL-1_TV	protists, planarian, nematodes, sea squirts	(Eickbush and Malik, 2002)
Hero	HEROTn, HERO-1_BF, HERO-1_SP	fish, sea urchin, lancelet	(Kojima et al., 2006)
Proto1	Proto1-1_NG, Proto-4_NG, Proto-6_NG	protist	(Kapitonov and Jurka, 2009a)
L1	DRE, Zorro, L1-40_XT, L1-11_XT, L1-1_XT, L1-1_DR, TDD3, Swimmer, L1-34_XT, L1-39_XT, L1, L1-38_XT, SHALINE14_MT, CIN4E_ZM, ATLINE1_1, SHALINE16_MT, Zepp, L1-1_CR, Ylli	plants, fungi, green algae, vertebrates, mammals	(Malik et al., 1999)
Tx1	KenoDr1, L1-56_XT, Tx1-1_NV, KenoFr1, L1-3_Cis, L1-55_XT, KoshiTn1, Tx1-2_NV, Tx1_XT	cnidarian, sea squirt, fish, frog	(Putnam et al., 2007)
Proto2	Proto2-1_HM, Proto2-2_HM, Proto2-1_SK, Proto2-1_BF, Proto2-1_CS1, Proto2-1_CS1, Proto2-3_CS1, Proto2-4_CS1, Proto2-5_CS1, Proto2-6_CS1, Proto2-7_CS1, Proto2-8_CS1	cnidarian, annelid, lancelet, acorn worm.	(Kapitonov and Jurka, 2009b)
RTETP	RTE-1_TP, RTETP-1_PM	diatoms	*
RTEX	RTEX-1_NV, RTEX-2_NV, RTEX-3_NV, RTEX-4_NV, RTEX-1_BF, RTEX-2_BF, RTEX-3_BF, RTEX-4_BF, RTEX-5_BF, RTEX-6_BF	cnidarians, lancelet	(Putnam et al., 2007)
RTE	RTE, Perere-3, RTE-1, SR2, Expander, Expander1_Cis, RTE1_ZM, RTE-14_BF, RTE-15_BF, RTE-1_BF, RTE-2_BF, SjR2, RTE-1_AG, RTE-1_DR, RTE-1_NV, RTE-3_NV, RTE-4_NV, RTE-5_NV	plants, cnidarians, insects, planarian, nematodes, lancelet, vertebrates, mammals	(Malik and Eickbush, 1998)
I	I_DM, IVK_DM, I-2_BM, Mosqu1_Aa2, I-1_DP, Loner	insects, crustaceans, lancelet, sea squirt, fish	(Malik et al., 1999)
Outcast	Outcast, Outcast-1_BF, Outcast-2_BF	mosquito, lancelet	(Biedler and Tu, 2003)
Nimb	I-1_BM, I-1_AA, I-3_DR, I-3_AC, I-4_AC, nimbus, I-1_SP, I-5_DR, I-1_CI, I-1_DR	insects, mollusks, fish	*
Ingi	Ingi-1_BF, Ingi-2_BF, I-2_AC, Ingi, Ingi2, I-1_AC	protists, mollusks, lancelet	(Eickbush and Malik, 2002)

Clade	Non-LTR retrotransposons	Host species	References
Jockey	FW, LINE-1_AA, Syrinx_DS, G5_DM, LDT1, Jockey, BMC1	insects, crustaceans	(Malik et al., 1999)
R1	RTAg4, R1_DM, R1, DMRT1A	fungi, insects, cnidarians	(Malik et al., 1999)
Loa	LOA, Baggins1_Cis, Baggins-2_NVi	insects, sea squirts	(Malik et al., 1999)
Tad1	I-1_AN, TRAS1, I-6_AO, Tad1, MGR583	fungi	(Malik et al., 1999)
Rex1	REX1_DR, REX1-4_XT, REX1-5_XT, CR1-9_NV, CR-10_NV, CR1-11_NV	cnidarians, vertebrates	(Voff et al., 2000)
CR1	T1, CR1-1a_XT, CR1-2_NV, CR1-3_NV, CR1-5_NV, CR1-6_BF, CR1-26_BF, DMC1A, CR1-2_AG, ZENON_BM, CR1-1_LG	cnidarians, insects, vertebrates	(Malik et al., 1999)
L2A	CR1-34_HM, L2-24_NV	cnidarians	*
L2B	CR1-1_AG, L2B-1_CP, L2B-1_HM	insects, cnidarians	*
L2	L2A, CR1-1_DR, CR1-2_DR, CR1-L2-1_XT, CR1-1_NV, CR1-16_NV, CR1-17_NV, CR1-3_Lme	cnidarians, insects, vertebrates	(Lovsin et al., 2001)
Daphne	Sake_BM, Daphne_DS, Daphne-1_BM, Daphne-1_TCa	crustaceans, insects	(Schon and Arkhipova, 2006)
Crack	Crack-1_CP, Crack-2_CP, Crack-3_CP, Crack-4_CP, Crack-1_NV, CR1-14_NV, CR1-12_SP, CR1-21_SP, Crack-1_BF, L2-2_Cis, L2-4_Cis, Crack-7_BF, Crack-24_BF, Crack-1_SP, Crack-1_CS1, CR1-7_HM, CR1-65_HM, Crack-1_IC	cnidarians, insects, sea urchin, lancelet, sea squirts	*

DNA and protein sequences of all listed non-LTR retrotransposons, including those that have not been reported in the literature, can be accessed from Repbase (Jurka et al., 2005) at <http://www.girinst.org/repbase/>.

* Reported in this manuscript.

Table 2

Mean values of errors δ of reclassification of the learning set by different phylogenetic methods

BIONJ	Clearcut	FastME	PHYLO_WIN	QuickTree	RaxML
3.9	4.4	4.4	6.2	4.6	6.7