Published online 17 September 2009

Nucleic Acids Research, 2009, Vol. 37, No. 21 ***e143***
doi:10.1093/nar/gkp752

MGEScan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes

Mina Rho[1] and Haixu Tang[1,2,*]

[1]School of Informatics and Computing, Indiana University, Bloomington, IN 47408 and [2]Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN 47404, USA

Received April 13, 2009; Revised August 13, 2009; Accepted August 27, 2009

ABSTRACT

Computational methods for genome-wide identification of mobile genetic elements (MGEs) have become increasingly necessary for both genome annotation and evolutionary studies. Non-long terminal repeat (non-LTR) retrotransposons are a class of MGEs that have been found in most eukaryotic genomes, sometimes in extremely high numbers. In this article, we present a computational tool, MGEScan-non-LTR, for the identification of non-LTR retrotransposons in genomic sequences, following a computational approach inspired by a generalized hidden Markov model (GHMM). Three different states represent two different protein domains and inter-domain linker regions encoded in the non-LTR retrotransposons, and their scores are evaluated by using profile hidden Markov models (for protein domains) and Gaussian Bayes classifiers (for linker regions), respectively. In order to classify the non-LTR retrotransposons into one of the 12 previously characterized clades using the same model, we defined separate states for different clades. MGEScan-non-LTR was tested on the genome sequences of four eukaryotic organisms, *Drosophila melanogaster*, *Daphnia pulex*, *Ciona intestinalis* and *Strongylocentrotus purpuratus*. For the *D. melanogaster* genome, MGEScan-non-LTR found all known 'full-length' elements and simultaneously classified them into the clades CR1, I, Jockey, LOA and R1. Notably, for the *D. pulex* genome, in which no non-LTR retrotransposon has been annotated, MGEScan-non-LTR found a significantly larger number of elements than did RepeatMasker, using the current version of the RepBase Update library. We also identified novel elements in the other two genomes, which have only been partially studied for non-LTR retrotransposons.

INTRODUCTION

Non-long terminal repeat (non-LTR) retrotransposons are a class of 'mobile genetic elements' (MGE, also called 'transposable elements', TEs) that are found in most eukaryotic genomes. They are often further classified as autonomous and non-autonomous, depending on whether or not a particular insertion encodes all the proteins needed for its own transposition. Reverse transcriptase (RT) is encoded by all autonomous non-LTR retrotransposons, and thus it has been used as the main signal to identify and classify these elements. Another important protein domain encoded by most non-LTR retrotransposons is the apurinic/apyrimidinic endonuclease (APE). Transposition of non-LTR retrotransposons requires the APE to cleave the chromosomal DNA at the target site. RT uses the 3′-end of the DNA break as the primer to reverse transcribe the element's RNA transcript into the new site (1). A small number of non-LTR retrotransposons encode an RNaseH domain. In the clades CRE, R2 and R4, the endonuclease is a restriction-enzyme-like endonuclease (REL-endo) instead of an APE domain (2–4).

Previous studies have shown that non-LTR retrotransposons can be grouped into 11 'clades' based on RT phylogeny (2). These analyses also identified 11 regions of high conservation or identity in RT protein domain sequences, by using a sliding-window similarity index. In addition to the 11 clades that were initially classified, several clades, including L2, NeSL-1 and Rex1, have recently been identified (5–9).

The MGEs, in particular the non-LTR retrotransposons, and their derived sequences constitute a large portion of eukaryotic genomes [e.g. at least 20% of the human genome is derived from non-LTR retrotransposons (10)]. Therefore, genome-wide identification and

*To whom correspondence should be addressed. Tel: +1 812 856 1859; Fax: +1 812 856 4764; Email: hatang@indiana.edu

classification is necessary for a better understanding of genome evolution, including changes in total genome size and architecture. Within this context, computational methods have been developed to identify MGEs in genomic sequences. One of the approaches widely applied is pair-wise sequence similarity-based search by using a library of known MGE sequences. RepeatMasker is one of the programs for this purpose, and has been extensively used for the annotation of newly sequenced genomes, to comprehensively identify and classify TEs. While this approach has a simple computational framework, which totally rely on the similarity on DNA sequences, it shows low sensitivity for the identification of novel MGEs, because of their low-sequence similarities with known elements. Some other methods utilizing the comparison of protein coding sequences [e.g. by using tblastn (11)], have been applied to find genomic fragments encoding RT domains (12). Even though these methods showed better performance than those based on reference DNA sequences, they require post-processing steps to merge and classify the protein matches. Finally, most of the pair-wise sequence similarity-based methods require a significant amount of computation time as the size of library increases, because they carry out pair-wise comparisons against each sequence in the library.

De novo approaches attempt to find repetitive sequences in a particular genome, and then use similarity comparisons to cluster TEs into repeat families (13,14). Since this approach relies on multiple copies of highly similar elements, it is difficult to identify MGE families where individual insertions have diverged from each other, or that have low-copy numbers in the target genome.

The other approach utilizes a combination of features obtained from the MGEs other than pairwise sequence similarities (15–19). For example, the methods for indentifying LTR retrotransposons (15,16) find a pair of LTRs by using the string pattern matching methods at the first step and locates the other features such as target site duplication and protein domains. These methods could detect better the insertion of entire long MGEs as a unit, compared with Repeatmasker or *de novo* approaches (13,14).

In this article, we develop a computational approach to the identification and classification of non-LTR retrotransposons in whole genomic sequences, using probabilistic models. Our overall computational architecture is inspired by the basic concept of a generalized hidden Markov model (GHMM) (20). As we described above, the most important features for identification in non-LTR retroelements are protein domains needed for their transposition. Protein domains are typically conserved around the some regions (e.g. the active sites), but the linker regions between these domains are far less conserved. In support of this observation, previous studies have confirmed that the 11 regions in reverse transcriptases are highly conserved (2). Under these circumstances, we considered that a profile HMM (pHMM) for an entire subsequence, combining RT, linker, and APE, might lower the overall sensitivity, primarily due to the lower sequence similarity of the linker region. Thus, we combine two seperate pHMMs for each domain, along with inter-domain linker region. The inter-domain linker regions represent a structural rather than functional component in the folded protein, and therefore are less conserved than domains. As such, more general criteria should be used that rely less on the specific amino acid sequences but more on the overall structural pattern that can be quantified using parameterized and pre-calibrated physical properties. We therefore used the well-established hydrophobicity scale of 20 amino acids to evaluate whether a given amino acid sequence flanked by domains is likely an inter-domain linker. Modularity is another advantage of our approach, which allows for other method/function to be incorporated in future studies. Finally, our model can be used as a classification tool for non-LTR retrotransposons, which classifies each element into one of the known 'clades' (2).

MATERIALS AND METHODS

The dataset

The genomic sequences of *Drosophila melanogaster* (BDGP Release 5, dm3) (21), *Ciona intestinalis* (ci2) (22) and *Strongylocentrotus purpuratus* (strPur2) (23) were downloaded from the UCSC Genome Browser web site (http://genome.ucsc.edu). The genomic sequence of *Daphnia pulux* (Release 1, JGI060905) were downloaded from the wfleaBase Daphnia Genome project web site (http://wfleabase.org). For comparison, non-LTR retrotransposons for the first three genomes, annotated using RepeatMasker, were downloaded from the UCSC Genome Browser web site (http://genome.ucsc.edu). In order to train the pHMM, RT and APE sequences (ds36752 and ds36736) used in the previous study (2) were collected from EMBL-EBI (http://www.ebi.ac.uk/webin-align/webin_align_listali.htm). The annotated non-LTR retrotransposons collected from RepBase Update (http://www.girinst.org/repbase/index.html) (RepBase13.07) were used as the training sequences for the overall model. The RT sequences used in the phylogenetic analysis were obtained from Genbank: *Takifugu* (AAD19348), *Paralichthys* (AAN15747), *Danio* (BAE46429), *Trimeresurus* (D31777), *Xiphophorus* (AF278692), *Oryzias* (AB054295), R2Ci-A-C (AB097121-3), HEROFr (AB097130), HEROTn (AB097131), HERODr (AB097132), YURECi (AB097133), EhRLE2 (AB097128), EhRLE3 (AB097129) and DongAg (AB097127). Additional RT sequences (ALIGN_000231 and 000232) used in the previous study (24) were collected from EMBL-EBI.

The model architecture

In order to model common features of non-LTR retrotransposons in a large variety of genomes, we built a model consisting of 12 super states, each corresponding to a different clade (Figure 1). Considering the current annotations and resolution in the phylogeny of non-LTR elements (2), we used two super states to represent two groups of closely related clades, one state (denoted

as R1) for the group of R1 and LOA clades, and the other state (denoted as R2) for the group of R2, R4, GENIE and NeSL clades. Because the clades LOA, R4, GENIE and NeSL contain very few known elements and thus are difficult to model separately, we combined them with the closest clades containing a sufficient number of known elements (the clades R1 and R2), into single super states. The other 10 super states correspond to single clades of non-LTR retrotransposons, including the L1, L2, RTE, Tad1, CRE, Jockey, CR1, I, Rex1 and RandI clades. Each super state consists of various numbers (numbers 1–3) of states, corresponding to protein domains and linker regions encoded by the non-LTR elements. The key components in our model can be summarized as follows.

(1) The states representing the RT domains, the APE domains and the linker regions encoded by the elements in each clade, and a separate state representing the genomic regions outside the non-LTR elements.
(2) The length distribution associated with each state. Since the length distributions of the same domains in different clades are similar, we use the same length distribution for the same domain (RT or APE) in all clades.
(3) The scores for each state. The scores from the states of protein domains are modeled by a pHMM, while the linker regions are modeled by Gaussian Bayes classifiers.

Searching non-LTR elements can then be expressed as a parsing problem by using dynamic programming approach. In order to evaluate the scores for the state of protein domains and inter-domain region, we adopted two probabilistic models, a pHMM (for the protein domains) and a Gussian Bayes classifier (for the linker regions). These models will be described in detail in the next sections. In general, the computational complexity of 'viterbi'-like parsing algorithm allowing variable length of segments is $O(Q^2L^2)$, in which Q is the number of states and L is the sequence length. This is practically infeasible for our problem when the genome sequence is very long. To reduce the complexity to $O(Q^2L)$, we fix the length of sequences in a state by locating potential signals for the start and end site of the protein domains in a pre-processing step.

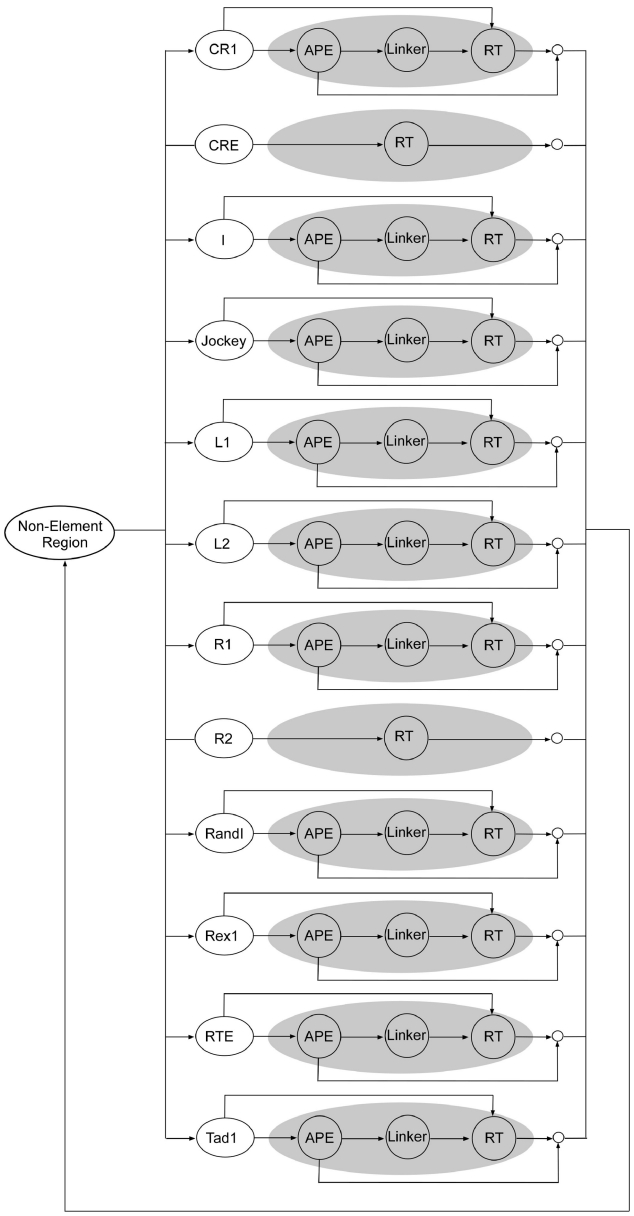

Figure 1. The model for identifying and classifying non-LTR retrotransposons. The circles represent states corresponding to protein domains and linker regions. The shaded ovals represent super states corresponding to clades. This model classifies the elements in 12 clades independently.

pHMMs for domains

Most autonomous elements of non-LTR retrotransposons encode more than one protein domain essential for transposition. Among the 12 clades considered here, elements in 10 clades encode both APE and RT domains in a single ORF, whereas elements in the clades CRE and R2 encode only a RT domain (Figure 1). Previous studies have shown that these domains are conserved among non-LTR retrotransposons in each clade and across the clades (2). Although the domains in different genomes show a range of sequence similarities, 11 regions are well conserved across many elements from different organisms. Eight of them are in the catalytic finger/palm region and three of them is in the thumb region of the 'right hand' structure (2,25). In addition, the positions of the conserved amino acids and the degree of their conservation show different patterns in different clades. Therefore, pHMM are a straightforward approach to representing these conservation patterns in the protein domain sequences (26,27). We applied different pHMMs to compute the score for the states of protein domains in each clade, which is the logarithm of the E-value from the pHMM divided by the logarithm of estimated minimum E-value (i.e. a constant value of $1.0E-300$).

Table 1. The number of non-LTR retrotransposons used in the genome-specific cross-validation

Clade	*D. melanogaster*	*D. rerio*	*S. purpuratus*
CR1	3 (71)	0 (74)	2 (72)
CRE	0 (4)	0 (4)	0 (4)
I	2 (7)	1 (8)	0 (9)
Jockey	23 (6)	0 (29)	0 (29)
L1	0 (203)	10 (193)	1 (202)
L2	0 (57)	4 (53)	16 (41)
R1	5 (21)	0 (26)	0 (26)
R2	2 (25)	0 (27)	0 (27)
RandI	0 (6)	0 (6)	0 (6)
Rex	0 (7)	1 (6)	0 (7)
RTE	0 (14)	1 (13)	0 (14)
Tad1	0 (3)	0 (3)	0 (3)

The numbers in each column denote the number of elements used for cross-validation. The numbers of elements from the other genomes used in the training are shown in parentheses.

We adopted an iterative method to build the pHMMs. Initially, we built two unified pHMMs for RT and APE domains, by using well-studied protein domain sequences as training sequences (2). Note that these models are common to the respective protein domains (RT or APE) of non-LTR retrotransposons in all clades. In the subsequent step, we used these pHMMs to search against all non-LTR retrotransposons in the RepBase Update library and identified the RT and APE sequences of non-LTR elements. In order to validate the clades that are assigned to these elements in RepBase Update, we built neighbor-joining trees with a Poisson correction model for all these domain sequences, along with those from the previously characterized reference elements (2). After eliminating 12 misclassified elements, a total of 470 RT sequences and 331 APE sequences were retained in order to train models. The number of elements from each of the 12 clades was shown in Supplementary Table S1. We carried out multiple alignments of the domain sequences in each clade using ClustalW (28) with the default parameter setting. The final pHMM models were then trained by using 'hmmbuild' in the HMMER package, based on the multiple alignments (27).

To validate the performance of the pHMMs in classification, we designed a genome-specific cross-validation method. Instead of partitioning the whole set of elements into subsets and using one of them at a time for the validation, we used the set of elements identified in a specific genome as the validation data and used the remaining elements to train the pHMMs. Since our main goal is to identify novel elements in the newly sequenced genomes, a genome-specific validation is desirable. We carried out the validations using the three phylogenetically distant *D. melanogaster*, *D. rerio* and *S. purpuratus* genomes (Table 1). For the *D. melanogaster* genome, elements in other *Drosophila* genomes were also excluded in the training set, because many elements are shared across *Drosophila* genomes. In these three independent cross-validation experiments, all the non-LTR elements were correctly classified into the annotated clade, thus reaching 100% accuracy. However, this might be explained by the fact that the annotated set consists of elements containing well-conserved protein domains. Additionally, we investigated the distribution of *E*-values among all putative clades reported by HMMER, for each of the domains in the annotated elements. A logarithmic plot of *E*-value distribution for the RT domains in the annotated non-LTR elements from the *Drosophila* genome is shown in Figure 2. In each vertical position (representing a separate element's domain), the highest point corresponds to the lowest *E*-value received for a particular clade, whereas the lower points correspond to the *E*-values received for the other 11 clades. In all cases, the highest point represents the annotated clade, and all the elements in the training set show clear differences in the *E*-values between the annotated clade and other clades, except for two elements (26 and 30). Plots constructed from the genomes of *D. rerio* and *S. purpuratus* showed similar trends (data not shown). Based on these analyses, we computed an *ad hoc* significance of the assigned clade for a given element using Equation (1).

$$Q = \frac{E_{\text{assigned}}}{E_{\text{alternative}}}. \quad \mathbf{1}$$

where E_{assigned} is the *E*-value obtained from the assigned clade; $E_{\text{alternative}}$ is the *E*-value that is the smallest one among the rest of the clades. Thus, a lower Q implies that the assigned clade is more significant. Two elements in the training set are weakly assigned ($Q > 10^{-3}$), whereas the other elements are reliably assigned ($Q < 10^{-10}$) (Supplementary Table S2). Our program reports the Q-value for all identified elements, to provide significance estimations for the predicted clade.

Classifiers for the linker regions

Non-LTR retrotransposons in 10 (out of 12) clades encode both RT and APE domains in a single ORF (Figure 1). For these clades, we define a state to model the 'linker regions', i.e. the amino acid sequence between APE and RT domains, attempting to build a predictor to compute the score for each of these regions that will distinguish the linker region encoded by an element from the linker regions encoded by the elements of the other clades, as well as arbitrary protein sequences. We observed that hydrophilic amino acid residues occur much more frequently than hydrophobic residues in the linker regions. This is consistent with previous studies showing that the linker regions between two active domains are often solvent-accessible and therefore contain more hydrophilic residues. In order to capture this sequence feature, we computed three hydrophobicity indices for the linker regions in known non-LTR retrotransposons. The classical Kyte and Doolittle (KD) scale (29) is derived from the changes in free energy upon transferring individual amino acids from water to a non-interacting isotropic phase. The partitioning of small model peptides between phospholipid membranes and buffer form the basis of the Wimley and White (WW) scale (30). A more biologically relevant hydrophobicity index, proposed by Hessa and von Heijne (HH) (31), was also used in our study. Since

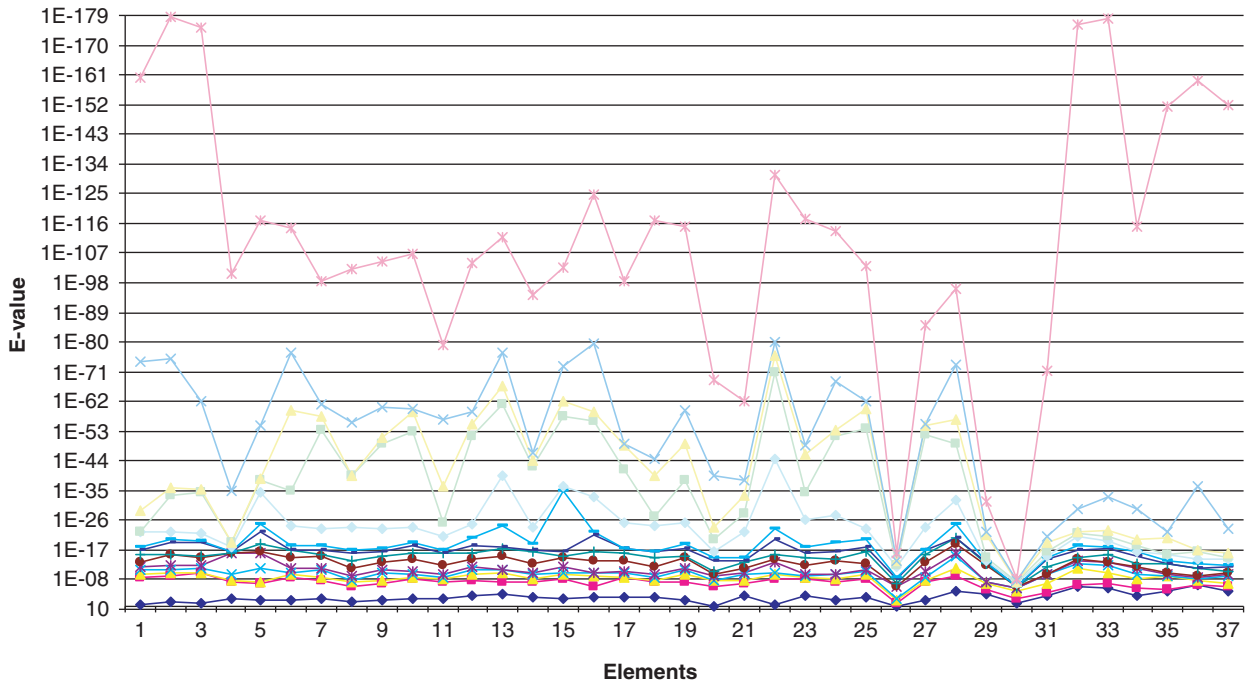

Figure 2. The distribution of *E*-values estimated by the pHMMs of 12 clades in *Drosophila* genomes. The *x*-axis represents each element (listed in the Supplementary Table S2) and the *y*-axis represents *E*-values. For each element, the *E*-values measured by 12 clades pHMMs are plotted. Each element was assigned to the clade from which the lowest *E*-value was obtained. Note that the highest point in each element (*x*-axis) does not mean that they belong to the same clade. Lines connecting each point were added for visualization purpose.

lower values in KD and WW scales correspond to increasing hydrophilicity, we negated the values in these two indices to make all values have consistent representation, i.e. the higher value indicates higher hydrophilicity. We obtained the APE, RT and linker regions separately from 325 elements in 10 clades in the training data set (Supplementary Table S1) and calculated the average values from the KD, WW and HH scales. The average KD values of the linker and domain regions are compared in Supplementary Figure S1. The values of the linker regions are consistently higher than those from both of the domain regions in 321 (out of 325) elements, which correspond to >99% of the training dataset. The other two indices, WW and HH, showed similar trends. We subsequently generated the distribution of the average hydrophobicity values both in the linker region and in a random model. A consistent and statistically distinctive distribution of the average hydrophobicity values was observed when comparing all three scales for the linker regions and random sequences (Figure 3). Three Gaussian models were trained using each of these indices, respectively, and then combined to calculate the probability that the region evaluated is a real linker region or a random protein sequence using Equations (2) and (3). The score of a sub-sequence S_i is the likelihood that S_i [with average hydrophobicity index $h(S_i)$] is a linker region, i.e. $h(S_i)$ is sampled from the distribution of average hydrophobicity index of linker regions, versus a random sequence (Equation 2). Thus, this predictor evaluates the probability that a subsequence S_i is from the linker region. The distribution of average hydrophobicity index of linker regions was fitted into a Gaussian distribution in Equation (3).

$$P(L_h|h(S_i)) = \frac{P(L_h)P(h(S_i)|L_h)}{\sum_{C\in\{L_h,R_h\}} P(C)P(h(S_i)|C)} \quad \mathbf{2}$$

$$P(h(S_i)|L_h) = \frac{1}{\sqrt{2\pi\sigma_{L_h}^2}} e^{\frac{-\left(h(S_i)-\mu_{L_h}\right)^2}{2\sigma_{L_h}^2}} \quad \mathbf{3}$$

where L_h is a model for linker regions analyzed by a hydrophobicity index h, which is one of {KD, WW, HH}, and R_h is a model for random sequences. The maximum probability among the three ones computed using three different hydrophobicity indices was defined as the final emission probability for the hidden states of the linker region (Figure 3).

Length distribution of states

The lengths of domains and linker regions were obtained from the same set of 470 elements that was used to obtain the scores for the domains. With a bin size of 100 bp, all the length distributions were displayed approximately as normal distributions (Supplementary Figure S2), which showed that 97% of the elements encode an APE domain with lengths from 600 to 800 bp, whereas 91% of the elements encode a RT domain with lengths from 1200 to 1400 bp.

Table 2. The number of 'ORF-preserving' non-LTR retrotransposons identified in the genomes of *D. melanogaster*, *D. pulex*, *S.purpuratus* and *C. intestinalis*

Clades[a]	*D. pulex*	*S. purpuratus*	*C. intestinalis*	*D. melanogaster*
R2[b]	–	–	11	–
R4[b]	–	–	1	–
NeSL[b]	27	37	5	–
L1	12	133	36	–
RTE	–	86	1	–
LOA[b]	23	–	4	3
R1[b]	–	–	–	18
Jockey	–	–	–	143
CR1	–	183	–	2
I	25	61	4	13
L2	49	489	–	–
Rex1	–	19	–	–
Total	136	1008	62	179

[a]No elements was found in three clades, Tad1, RandI and CRE.
[b]Elements in R1 and R2 groups were further classified by using phylogenetic analysis.

Implementation

We implemented a software tool named MGEScan-non-LTR in three modules. The first module prepares the input for evaluation process by locating the signals of domains in a given genomic sequence. The second module finds the optimal path of states, which corresponds to the annotation of the protein domains and linker regions in the clades. The third module post-processes the results and reports the coordinates and sequences of non-LTR retrotransposons classified into specific clades. The second (main) module was implemented in C and the other two modules were implemented in Perl. We used ClustalW (32) and HMMER (27) to build pHMMs for each domain. Since our method focuses on the identification of elements which retain the protein domains, the results include elements containing all the domains in ORFs, and those containing one of the domains but with partially truncated 5′- or 3′-ends. MGEScan-non-LTR can be freely downloaded at the Supplementary web site (http://darwin.informatics.indiana.edu/cgi-bin/evolution/nonltr/nonltr.pl).

RESULTS

We applied MGEScan-non-LTR to the complete eukaryotic genomes of *D. melanogaster*, *D. pulex*, *S. purpuratus* and *C. intestinalis*. In order to evaluate the sensitivity of the method, we first applied MGEScan-non-LTR to the *D. melanogaster* genome, which has been extensively studied for TEs in the entire genomic sequence. The *D. pulex* genome has recently been sequenced and no non-LTR retrotransposon had previously been annotated. This offers the opportunity to use MGEScan-non-LTR to investigate the diversity of elements in a newly sequenced genome. Even though many elements have been identified in the *S. purpuratus* and *C. intestinalis* genomes by previous studies (5,33), we were able to identify novel elements from several different clades. In comparison to RepeatMasker, we observe significantly more elements identified by MGEScan-non-LTR. We also present phylogenetic analysis of the elements identified in this study. The novel elements identified in this study were deposited in Genbank (accession numbers FJ905841-6).

Table 3. Summary of running time on the genomes of *D. melanogaster*, *D. pulex*, *S.purpuratus* and *C. intestinalis*

Genome	Length (MB)	Number of scaffolds (chromosomes)	Running time (h:min)
C. intestinalis	173	4390	2:18
D. pulex	227	9079	2:48
D. melanogaster	120	6	1:20
S. purpuratus	907	114 223	28:00

Performance evaluation

In *D. melanogaster*, a total of 179 'ORF-conserving' elements were identified from the five clades of CR1, I, Jockey, LOA and R1 (Table 2). In good agreement with previous studies (34,35), 'Jockey' is the most abundant clade among these five. For comparison, we collected previously known non-LTR retrotransposons from the Berkeley Drosophila Genome Project site and used RepeatMasker to map them to the genome. MGEScan-non-LTR identified all of the 107 'full-length' elements that were previously identified (>97% of the canonical element, excluding non-autonomous) (35). The other 72 identified elements overlapped with the 'fragmented' elements reported by RepeatMasker. This result indicates that there is no mis-identified element (false positive). Even though the currently annotated elements in the training set might be biased for *Drosophila* genomes, the genome-specific cross-validation that we conducted in the training process confirmed that MGEScan-non-LTR can successfully identify and classify novel elements.

MGEScan-non-LTR is fast enough for genome-wide annotation. The running time was measured for the entire process for the given genomic sequences (Table 3). The experiment was performed on an Intel xeon CPU 2 GHz running on Red Hat 4.1.2. For the *D. melanogaster*, *D. pulex* and *C. intestinalis* genomes, the running time was nearly linear to their genome sizes. In the case of *S. purpuratus* genome, however, the running time deviated from this general trend. The longer running time observed in this case might be explained by the fact that the large number of scaffolds slowed down the program since MGEScan needs extra cost for significantly larger amount file I/Os each scaffold file.

Non-LTR retrotransposons in the *D. pulex* genome

We identified 136 'ORF-preserving' non-LTR retrotransposons in the *D. pulex* genome, which were classified in the five clades I, L1, L2, LOA and NeSL. In contrast to other arthropod genomes, the most abundant clade in the *D. pulex* genome is the L2 clade, which has mainly been identified in *Deuterostomia* genomes, such as vertebrates and echinoderms (5) (Table 2). Interestingly, the *D. pulex*

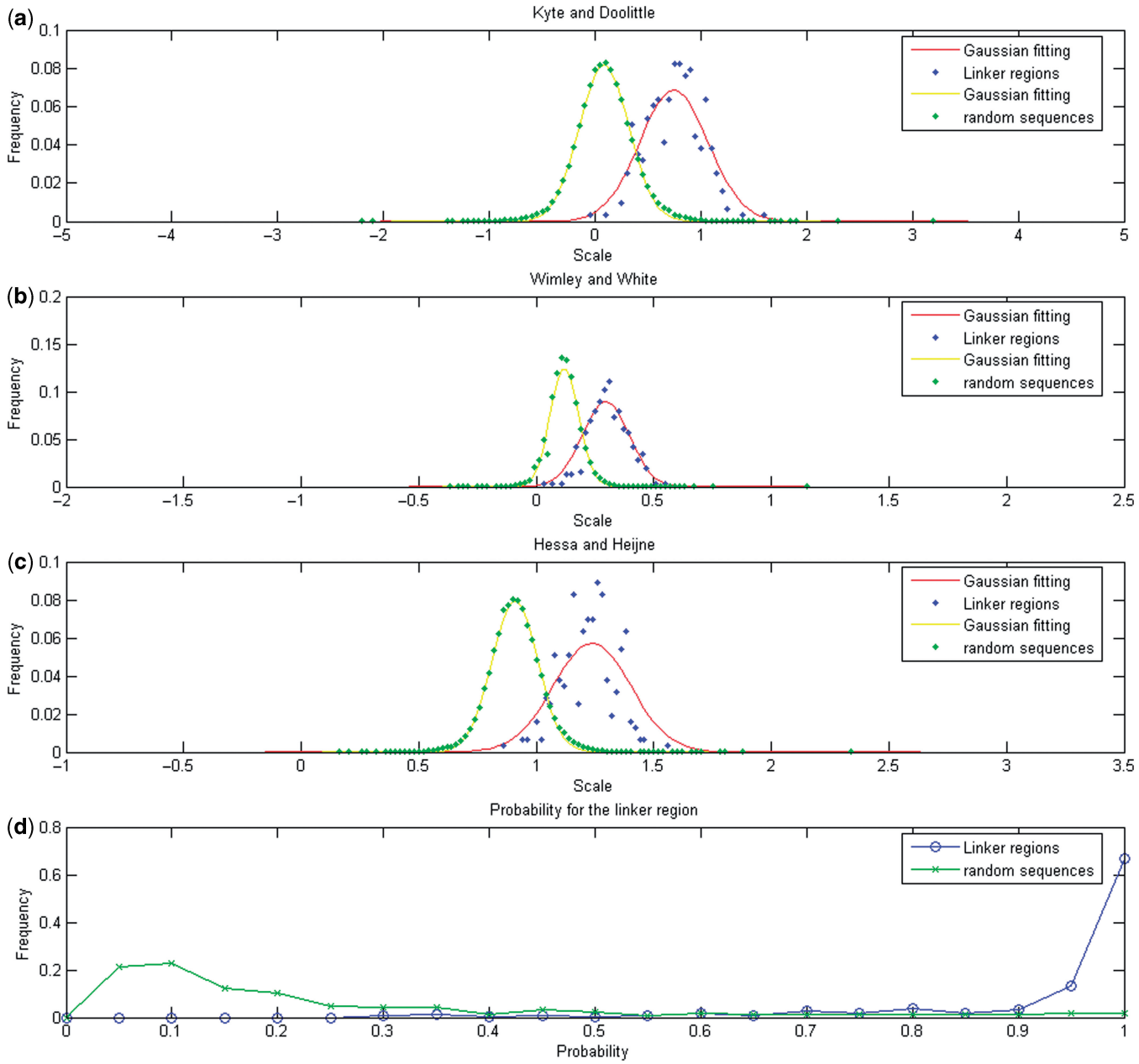

Figure 3. (**a**) The distribution of hydrophobicity values in KD, (**b**) WW and (**c**) HH hydrophobicity scales, and (**d**) probability for linker region. In (a)–(c), the *x*-axis represents the hydrophobicity scale and the *y*-axis represents the frequency. The distribution on the left-hand side with green dots was plotted with the values obtained from random sequences. The distribution on the right-hand side with blue dots was plotted with the values obtained from the elements in the training set. The yellow and red lines represent Gaussian distribution fitting of the data plotted. In (d), the *x*-axis represents the probability for the linker region and the *y*-axis represents the frequency.

L2 elements are closely clustered with the L2 elements from fish genomes such as *Takifugu rubripes*, *Oryzias latipes* and *Danio rerio* (Figure 4). The two representative elements in insect genomes *Bombyx mori* and *Aedes agypti* are clustered each other, but are further from the *D. pulex* L2 elements than the fish elements. In addition, a high level of conservation in RT protein sequences (~56%) between the L2 elements in the *D. pulex* genome and the elements in *Takifugu* (>700 MYA apart) was observed.

Of the LOA clade, 22 elements were identified in the *D. pulex* genome. The LOA elements have mostly been identified in the *Drosophila* (LOA, TRIM, and Bilbo) (36,37), *Aedes* (Lian) (38) and *Ciona* (CiLOA) (33) genomes. Phylogenetic analysis of the LOA elements showed a strong relationship with known LOA elements (boostrap value > 65%) (Figure 4). However, while all known LOA elements (LOA, TRIM, Bilbo, Lian and CiLOA) contain an RNase H domain downstream of the RT domain, no evidence was observed for the

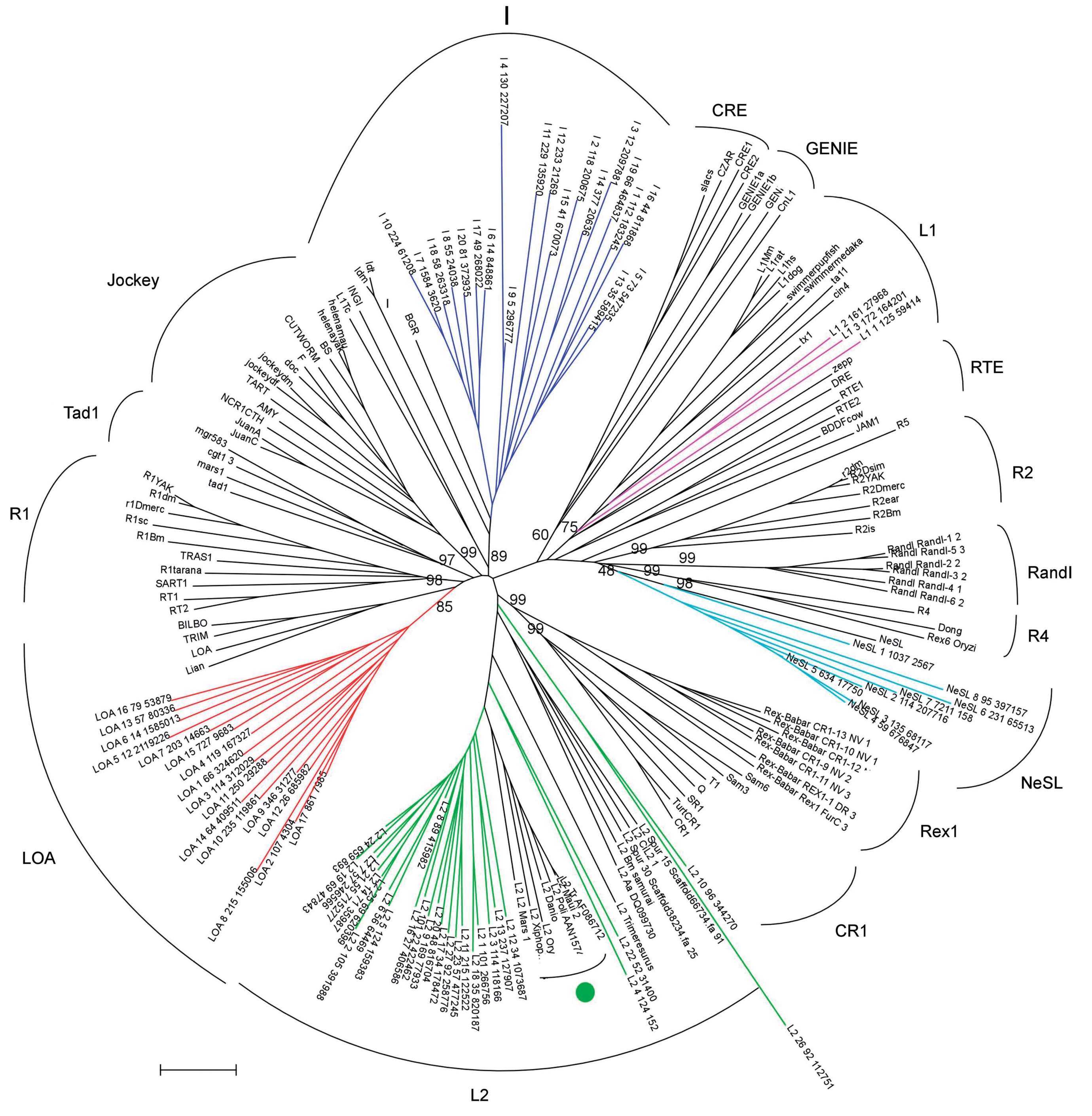

Figure 4. Phylogenetic analysis of RT domain sequences in the *D. pulex* elements, along with previously known elements from other genomes. For the *D. pulex* elements, LOA elements are highlighted in red, I elements in blue, L2 elements in green, L1 elements in pink and NeSL elements in cyan. A cluster of L2 elements (L2_Tr_AF086712, L2_Maul_2, L2_Poll_AAN15747, L2 Danio, L2_Oryzlas, L2_Xlphophorus and L2_L2_Mars_1) from fish genomes were indicated by green circle.

domain in the *D. pulex* elements by BLAST and Pfam domain search (against PF00075).

Even though we could not generate separate models for the ancient clades CRE, GENIE (24), R2 (3), R4 and NeSL (6) because there are too few known members, (4) (see 'Materials and Methods' section), we conducted phylogenetic analysis of the elements identified as in the R2 group, in order to obtain further resolution of their classification. A neighbor-joining tree with bootstrapping was constructed with the intact RT sequences from the five ancient elements, along with RT sequences from known elements of other organisms. As shown in Figure 5, the elements *Dpul_a1-5* cluster with the elements HEROFr, HEROTn and HERODr in the

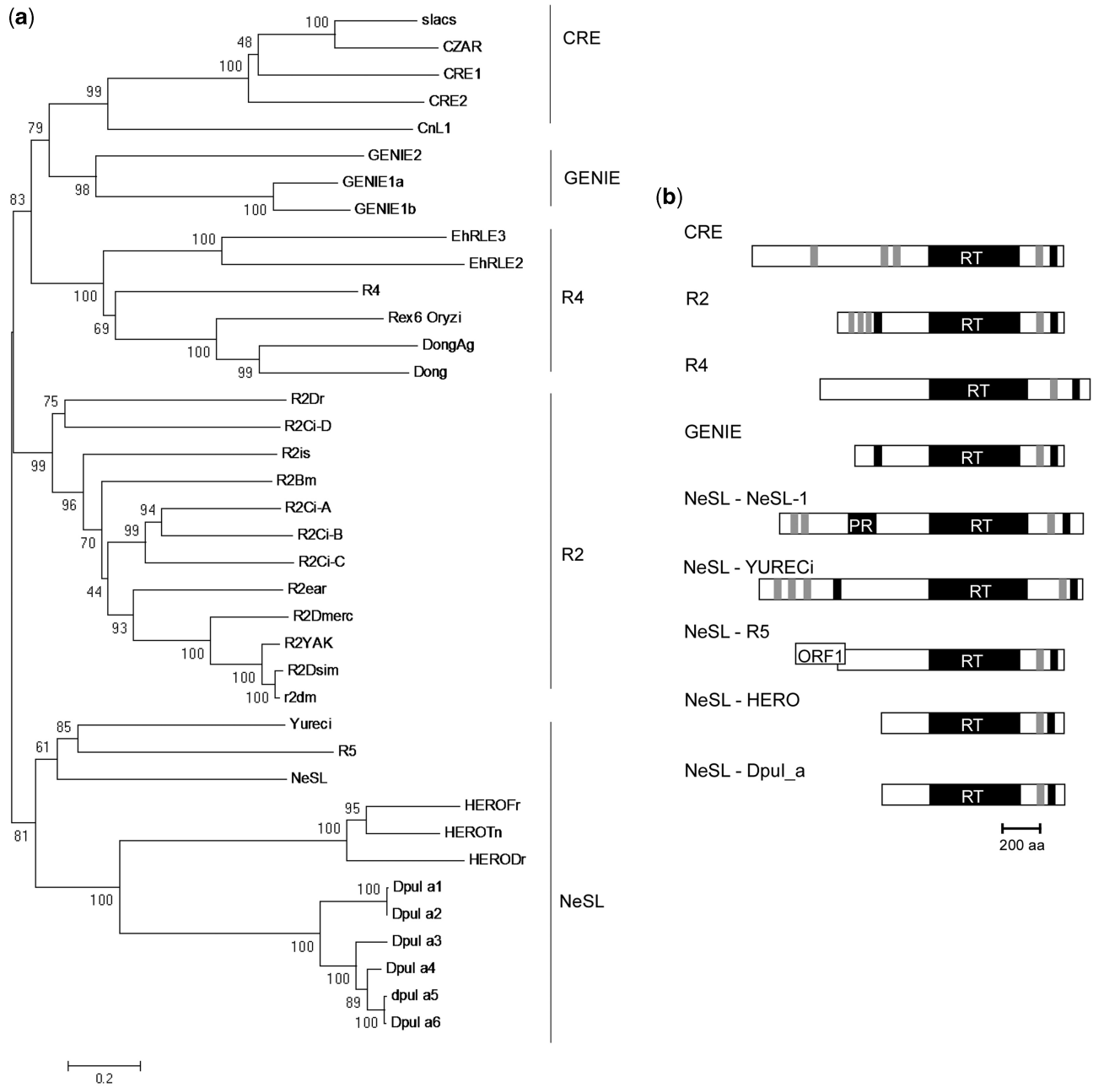

Figure 5. (**a**) Phylogenetic tree and (**b**) schematic representation of the elements in five ancient clades of CRE, GENIE, R2, R4 and NeSL. The neighbor-joining tree was built with 1000 rounds of bootstrapping. Note that the cluster of *D. pulex* elements is connected with the cluster of HEROs in the NeSL clades with a very high-bootstrap value (100%). The schematic representation of the *D. pulex* NeSL elements is most similar with one of the HERO elements. The gray bars downstream of the RT are cysteine−histidine motif (C-X2-C-X8-H-X4-C). The blank bars downstream of RT represent the REL domain motif (PD..D).

NeSL clade with very high-boostrap value (100%), although with quite deep branches. The cysteine-histidine motif (C-X2-C-X8-H-X4-C) and the REL domains motif (PD..D), which are downstream of the RT domains, are observed in all five elements (Figure 5b). This is a common feature of the ancient clades (6). However, upstream of the RT domains, no motifs were observed in either HERO or Dpul_a elements, while the other ancient clades had cysteine−histidine motifs upstream of the RT domain (39). Given the definition of 'clade' (2), it might not be appropriate to combine all five elements (Yureci, R5, NeSL, HERO and *Dpul_a*) into the same clade, since they have different structural features (Figure 5). However, considering that a limited number of elements have been identified to date, the classification can only be refined after more elements have been identified.

In order to compare the performance of MGEScan-non-LTR with RepeatMasker, we ran RepeatMasker on the *D. pulex* genome using the current version of Repbase Update library. This is an interesting comparison, since

MGEScan-non-LTR used the same library as the training set for its model. RepeatMasker identified 19 elements (length >1000 bp), compared to our 136 elements, and 15 of them overlapped with our results. This result clearly shows that our method performs better than homology-based searches in identifying novel elements. Interestingly, all the elements identified by RepeatMasker were classified as L2 clade. This result is consistent with the high level of sequence conservation among L2 elements.

Non-LTR retrotransposons in the *S. purpuratus* genome

We identified a total of 1008 non-LTR retrotransposons, which were classified into the seven clades CR1, I, L1, L2, NeSL, RTE and Rex1. To the best of our knowledge, this is the first genome-wide survey of non-LTR retrotransposons in the *S. purpuratus* genome, although elements in the CR1, L1, L2, NeSL and RTE clades have been identified by previous studies (39,40). Even though the number of 'ORF-preserving' non-LTR retrotransposons identified is much higher than those in other three genomes studied here, the density (1.11 elements/MB) is only slightly higher than in the *D. pulex* genome (1.06 elements/MB). The *Rex1* clade has mostly been found in fish genomes, such as *F. rubripes* and *T. nigroviridis* (7). Notably, we found 19 *Rex1* elements in the *S. purpuratus* genome and six of them contain intact RT sequences that have not been degraded by a stop codon or frameshift mutation. Phylogenetic analysis of these six elements with known elements, mostly from fish genomes, showed that the *S. purpuratus Rex1* elements are more closely related with *Babar* elements from *Battrachocottus baikalensis*, *Oncorhynchus keta* and *T. nigroviridis* (bootstrap value 97%) (Supplementary Figure S2). RepeatMasker identified two of them as in the CR1 clade, while four of them were not identified.

Non-LTR retrotransposons in the *C. intestinalis* genome

The *C. intestinalis* is an important model organism, as a chordate that predates the vertebrate lineage. Interestingly, the *C. intestinalis* genome has divergent elements from the clades that do not have APE domain: 11 R2 elements, one R4 element and five NeSL elements. In the phylogenetic analysis (Supplementary Figure S3), the NeSL elements are most closely related with the *D. pulex* elements identified in this study, with a boostrap value of 68%. However, the similarities of the elements between the genomes are low (<40%), suggesting that both lineages have evolved through vertical transmission. The similarities of RT sequences between the NeSL elements in the *C. intestinalis* genome were quite high (>86%); but all of them have suffered multiple stop codons. MGEScan-non-LTR also reported a novel element in the RTE clade. We call this element 'CiRTE', following the naming scheme used in the previous paper (33). A BLASTX search with 'CiRTE' as query showed the most significant hit (*E*-value: 2*e*–54) in the *S. purpuratus* genome. The RT sequence of 'CiRTE' was aligned with those of the previously annotated elements *CiI*, *CiL1*, *CiL2*, *CiLOA* and *CiR2* from the *C. intestinalis* genome to identify the highly conserved sequences of the RT domain, showing seven well-conserved regions (Figure 6). In phylogenetic analysis (Supplementary Figure S3), 'CiRTE' clustered with RTE elements in the Cow and Vipera genomes with a high bootstrap value (>94%).

DISCUSSION

With an increasing number of sequenced genomes available, efficient computational methods are required for genome-wide identification of MGEs, to assist both genome annotation and evolutionary studies. MGEScan-non-LTR can identify novel elements from the diverse clades of non-LTR retrotranposons from genomic sequences, and has three important advantages as described below.

First, it is fully automated. Given a genomic sequence, MGEScan-non-LTR will report as the final output all non-LTR retrotransposons that are identified and classified to specific clades. Due to the mechanism of the reverse transcription, 5′ truncation occurs frequently when new elements were inserted into the host genome. As a result, it is difficult to find 'full-length' non-LTR retrotransposons, although the structure of ORFs encoding protein domains may be well-preserved in many elements. Our method mainly focuses on the identification of non-LTR retrotransposons preserving all the required domain regions for transposition. These 'ORF-preserving' elements identified in the entire genomic sequences serve as a basic set to analyze MGEs. After they are identified, fragmented or truncated copies can be found with a RepeatMasker search, using these 'ORF-preserving' elements as the library.

Second, MGEScan-non-LTR is modular and extendible. It assigns each essential segment of non-LTR retrotransposons to a state that has an independent model for calculating scores. As such, these models can be modified or replaced by better probabilistic models, and new models can be incorporated for more diverged elements (clades) in the future. Our method has a more generalized architecture that can have separate methods or functions to capture the best features associated with the specific region. Note that the inter-domain linker regions need to be analyzed using more general criteria that do not rely on sequence homology, we used parameterized and pre-calibrated physical properties such as hydrophobicity of amino acid residues.

Third, MGEScan-non-LTR scales well with increasing element types and numbers. Models are built upon the best knowledge and data that are presently available. When the first element in a new clade is encountered, MGEScan might not be able to unambiguously differentiate it from those in the closest clade. Our approach to addressing this issue is to devise a method that allows one to incorporate a new clade. After a number of elements cluster together with high-bootstrap value in phylogeny and have the same structure of domains, one could suggest a new clade for this putative new group. Our current model can easily incorporate a newly defined clade by adding new superstates. Extendibility to handle

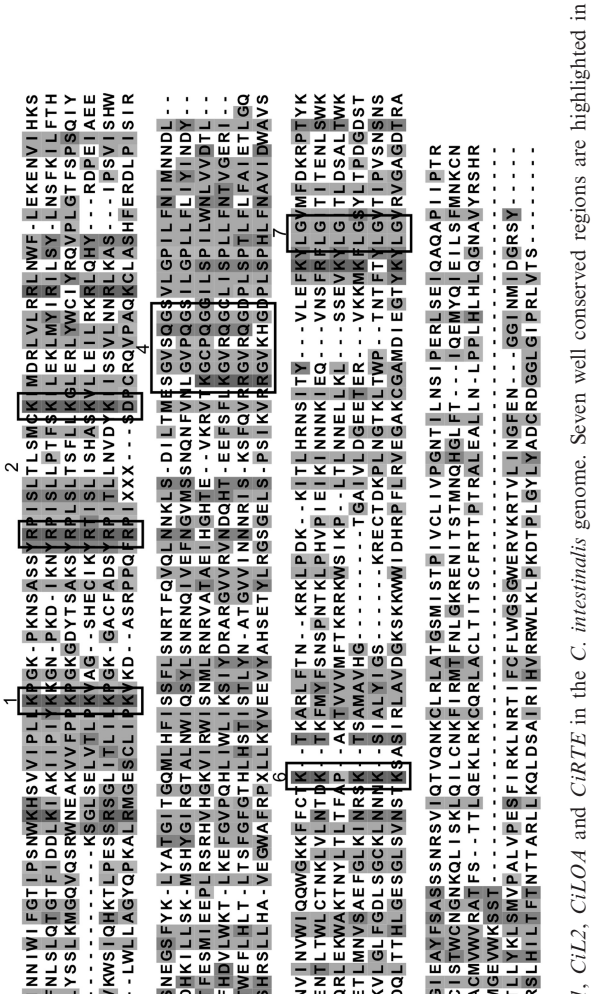

Figure 6. Multiple sequence alignments of the RT domain sequences from *CiI, CiL1, CiL2, CiLOA* and *CiRTE* in the *C. intestinalis* genome. Seven well conserved regions are highlighted in boxes.

such situation is one of the important advantages of our model.

In addition, the increased number of elements does not increase searching time significantly since all the elements in the training set are not compared with the genome individually, but are used to train the model, which is subsequently used for the genome-wide search. This feature will increasingly give MGEScan-non-LTR an advantage, as more elements become available. Given these advantages, MGEScan-non-LTR should facilitate further systematic analysis of non-LTR retrotranposons in eukaryotic genomes. Based on the results from four different genomes in this initial study, we showed that MGEScan-non-LTR is ready to be used for the identification and classification of novel non-LTR retrotransposons.

ACCESSION NUMBERS

FJ905841–FJ905846

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Prof Michael Lynch, Dr Thomas Doak and Kwangmin Choi for helpful discussions. We are grateful to the Daphnia genome consortium for allowing us to access the complete genome sequence of *D. pulex* before its publication.

FUNDING

MetaCyt Initiative at Indiana University (funded by Lilly Endowment, Inc.). Funding for open access charge: MetaCyt Initiative at Indiana University (funded by Lilly Endowment, Inc.)

Conflict of interest statement. None declared.

REFERENCES

1. Luan,D.D., Korman,M.H., Jakubczak,J.L. and Eickbush,T.H. (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell*, **72**, 595–605.
2. Malik,H.S., Burke,W.D. and Eickbush,T.H. (1999) The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.*, **16**, 793–805.
3. Burke,W.D., Malik,H.S., Jones,J.P. and Eickbush,T.H. (1999) The domain structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods. *Mol. Biol. Evol.*, **16**, 502–511.
4. Burke,W.D., Singh,D. and Eickbush,T.H. (2003) R5 retrotransposons insert into a family of infrequently transcribed 28S rRNA genes of planaria. *Mol. Biol. Evol.*, **20**, 1260–1270.
5. Lovsin,N., Gubensek,F. and Kordis,D. (2001) Evolutionary dynamics in a novel L2 clade of Non-LTR retrotransposons in deuterostomia. *Mol. Biol. Evol.*, **18**, 2213–2224.
6. Malik,H.S. and Eickbush,T.H. (2000) NeSL-1, an ancient lineage of site-specific Non-LTR retrotransposons from *Caenorhabditis elegans*. *Genetics*, **154**, 193–203.

7. Volff,J.-N., Korting,C. and Schartl,M. (2000) Multiple lineages of the Non-LTR retrotransposon Rex1 with varying success in invading fish genomes. *Mol. Biol. Evol.*, **17**, 1673–1684.
8. Hizer,S.E., Tamulis,W.G., Robertson,L.M. and Garcia,D.K. (2008) Evidence of multiple retrotransposons in two litopenaeid species. *Anim. Genet.*, **39**, 363–373.
9. Novikova,O., Fet,V. and Blinov,A. (2009) Non-LTR retrotransposons in fungi. *Funct. Integr. Genomics*, **9**, 27–42.
10. International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
11. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
12. Biedler,J. and Tu,Z. (2003) Non-LTR retrotransposons in the african malaria mosquito, anopheles gambiae: unprecedented diversity and evidence of recent activity. *Mol. Biol. Evol.*, **20**, 1811–1825.
13. Bao,Z. and Eddy,S.R. (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.*, **12**, 1269–1276.
14. Edgar,R.C. and Myers,E.W. (2005) PILER: identification and classification of genomic repeats. *Bioinformatics.*, **21**, i152–i158.
15. McCarthy,E.M. and McDonald,J.F. (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics*, **19**, 362–367.
16. Rho,M., Choi,J.-H., Kim,S., Lynch,M. and Tang,H. (2007) *De novo* identification of LTR retrotransposons in eukaryotic genomes. *BMC Genomics*, **8**, 1811–1825.
17. Sperber,G.O., Airola,T., Jern,P. and Blomberg,J. (2007) Automated recognition of retroviral sequecnes in genomic data. *Nucleic Acids Res.*, **35**, 4964–4976.
18. Peterson-Burch,B.D., Nettleton,D. and Voytas,D.F. (2004) Genomic neighborhoods for *Arabidopsis* retrotransposons: a role for targeted integration in the distribution of the Metaviridae. *Genome Biol.*, **5**.
19. McClure,M.A., Richardson,H.S., Clinton,R.A., Hepp,C.M., Crowther,B.A. and Donaldson,E.F. (2004) Automated characterization of potentially active retroid agents in the human genome. *Genomics*, **85**, 512–523.
20. Rabiner,L.R. (1989) A tutorial of hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.
21. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.
22. Dehal,P., Satou,Y., Campbell,R.K., Chapman,J., Degnan,B., De Tomaso,A., Davidson,B., Di Gregorio,A., Gelpke,M., Goodstein,D.M. *et al.* (2002) The draft genome of ciona intestinalis: insights into chordate and vertebrate origins. *Science*, **298**, 2157–2167.
23. Sea Urchin Genome Sequencing, Consortium, Sodergren,E., Weinstock,G.M., Davidson,E.H., Cameron,R.A., Gibbs,R.A., Angerer,R.C., Angerer,L.M., Arnone,M.I., Burgess,D.R. *et al.* (2006) The genome of the sea urchin strongylocentrotus purpuratus. *Science*, **314**, 941–952.
24. Burke,W.D., Malik,H.S., Rich,S.M. and Eickbush,T.H. (2002) Ancient lineages of non-LTR retrotransposons in the primitive eukaryote, giardia lamblia. *Mol. Biol. Evol.*, **19**, 619–630.
25. Unge,T., Knight,S., Bhikhabhai,R., Lovgren,S., Dauter,Z., Wilson,K. and Strandberg,B. (1994) 2.2 A resolution structure of the amino-terminal half of HIV-1 reverse transcriptase (fingers and palm subdomains). *Structure*, **2**, 953–961.
26. Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D. (1994) Hidden Markov models in computational biology: applications to protein Modeling. *J. Mol. Biol.*, **235**, 1501–1531.
27. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
28. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) ClustalW and ClustalX version 2.0. *Bioinformatics*, **23**, 2947–2948.
29. Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.
30. Wimley,W.C. and White,S.H. (1996) Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat. Struct. Biol.*, **3**, 842–848.
31. Hessa,T., Kim,H., Bihlmaier,K., Lundin,C., Boekel,J., Andersson,H., Nilsson,I., White,S.H. and von Heijne,G. (2005) Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature*, **433**, 377–381.
32. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
33. Permanyer,J., Gonazalez-Duarte,R. and Albalat,R. (2003) The non-LTR retrotransposons in *Ciona intestinalis*: new insights into the evolution of chordate genomes. *Genome Biol.*, **4**, R73.
34. Berezikov,E., Bucheton,A. and Busseau,I. (2000) A search for reverse transcriptase-coding sequences reveals new non-LTR retrotransposons in the genome of *Drosophila melanogaster*. *Genome Biol.*, **1**, research0011.0011–0015.
35. Kaminker,J.S., Bergman,C.M., Kronmiller,B., Carlson,J., Svirskas,R., Patel,S., Frise,E., Wheeler,D.A., Lewis,S.E., Rubin,G.M. *et al.* (2002) The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biol.*, **3**, research0084.0081–0020.
36. Blesa,D., Gandia,M. and Martinez-Sebastian,M.J. (2001) Distribution of the bilbo non-LTR retrotransposon in Drosophilidae and its evolution in the Drosophila obscura species group. *Mol. Biol. Evol.*, **18**, 585–592.
37. Blesa,D. and Martinez-Sebastian,M.J. (1997) Bilbo, a non-LTR retrotransposon of Drosophila subobscura: a clue to the evolution of LINE-like elements in Drosophila. *Mol. Biol. Evol.*, **14**, 1145–1153.
38. Tu,Z., Isoe,J. and Guzova,J.A. (1998) Structural, genomic, and phylogenetic analysis of lian, a novel family of non-LTR retrotransposons in the yellow fever mosquito, aedes aegypti. *Mol. Biol. Evol.*, **15**, 837–853.
39. Kojima,K.K. and Fujiwara,H. (2004) Cross-genome screening of novel sequence-specific non-LTR retrotransposons: various multicopy RNA genes and microsatellites are selected as targets. *Mol. Biol. Evol.*, **21**, 207–217.
40. Jurka,J., Kapitonov,V.V., Pavlicek,A., Klonowski,P., Kohany,O. and Walichiewicz,J. (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogen. Genome Res.*, **110**, 462–467.