Grechishnikova and Poptsova *BMC Genomics* (2016) 17:992
DOI 10.1186/s12864-016-3344-4

BMC Genomics

RESEARCH ARTICLE Open Access

Conserved 3′ UTR stem-loop structure in L1 and Alu transposons in human genome: possible role in retrotransposition

Daria Grechishnikova[1] and Maria Poptsova[1,2*]

Abstract

Background: In the process of retrotransposition LINEs use their own machinery for copying and inserting themselves into new genomic locations, while SINEs are parasitic and require the machinery of LINEs. The exact mechanism of how a LINE-encoded reverse transcriptase (RT) recognizes its own and SINE RNA remains unclear. However it was shown for the stringent-type LINEs that recognition of a stem-loop at the 3′UTR by RT is essential for retrotransposition. For the relaxed-type LINEs it is believed that the poly-A tail is a common recognition element between LINE and SINE RNA. However polyadenylation is a property of any messenger RNA, and how the LINE RT recognizes transposon and non-transposon RNAs remains an open question. It is likely that RNA secondary structures play an important role in RNA recognition by LINE encoded proteins.

Results: Here we selected a set of L1 and Alu elements from the human genome and investigated their sequences for the presence of position-specific stem-loop structures. We found highly conserved stem-loop positions at the 3′UTR. Comparative structural analyses of a human L1 3′UTR stem-loop showed a similarity to 3′UTR stem-loops of the stringent-type LINEs, which were experimentally shown to be recognized by LINE RT. The consensus stem-loop structure consists of 5–7 bp loop, 8–10 bp stem with a bulge at a distance of 4–6 bp from the loop. The results show that a stem loop with a bulge exists at the 3′-end of Alu. We also found conserved stem-loop positions at 5′UTR and at the end of ORF2 and discuss their possible role.

Conclusions: Here we presented an evidence for the presence of a highly conserved 3′UTR stem-loop structure in L1 and Alu retrotransposons in the human genome. Both stem-loops show structural similarity to the stem-loops of the stringent-type LINEs experimentally confirmed as essential for retrotransposition. Here we hypothesize that both L1 and Alu RNA are recognized by L1 RT via the 3′-end RNA stem-loop structure. Other conserved stem-loop positions in L1 suggest their possible functions in protein-RNA interactions but to date no experimental evidence has been reported.

Keywords: L1, Alu, LINE, SINE, Mechanisms of retrotransposition, Stem-loop, Stem-loop structures

Background

L1 family is by its mass the largest family of Long INterspersed Elements (LINEs) in humans; L1s are present in more than 500 000 copies and occupy approximately 17% of the total human genome length [1]. Most of L1s found in the human genome are 5′-end truncated or damaged and therefore have lost the retrotransposition ability [2]. It is estimated that, on average, there are 80–100 L1s in a human being, which are still able to move through the genome [3]. A typical L1 is about 6 kb in length and consists of a 5′ untranslated region (UTR), two non-overlapping open reading frames (ORFs), a 3′ UTR and a poly-A tail.

L1 transcription is initiated by an RNA polymerase II promoter located at the 5′ UTR [4, 5]. The first open reading frame (ORF1) encodes a 40 kDa protein (ORF1p) consisting of a coiled-coil domain [6], a non-canonical RNA recognition motif domain [7, 8] and a basic carboxyl-terminal domain [9]. Although the exact

* Correspondence: maria.poptsova@gmail.com
[1]Physics Department, Lomonosov Moscow State University, Moscow 119991, Russia
[2]National Research University Higher School of Economics, Moscow 101000, Russia

role of ORF1 is not clear, it was demonstrated that this protein is required for retrotransposition [10]. The second protein, ORF2, is a 150 kDa protein and it combines endonuclease (EN) [11] and reverse transcriptase (RT) [12] activities. The L1 3′UTR is about 200 bp and is poorly conserved. It contains a polypurine tract that potentially could form a G-quadruplex structure [13], but the function of this sequence remains unknown. L1 ends with a poly-A tail that was shown to be critical for L1 retrotransposition [14]. Short direct repeats, flanking a transposon, are generated from the target DNA sequence during the L1 integration. The length of repeats can range from a few to several hundred nucleotides [2].

The second family of retrotransposons by mass in the human genome is Alu, belonging to the class of Short Interspersed Elements (SINEs). By copy number Alu is the most ubiquitous retrotransposon with more than 1 mln copies present in the human genome and occupying approximately 11% of the total genome length [1]. The length of a typical Alu is around 300 bp. Though almost all SINEs are derived from tRNA [15], Alus are derived from 7SL RNA, which functions as a component of the signal recognition particle [16, 17]. It consists of two monomers, each similar to the 7SL RNA, an A-rich connecting domain and a poly-A tail varying in length. In contrast to L1 Alu cannot amplify by itself as it does not encode any proteins and hires the L1 retrotransposition machinery [18, 19].

The mechanism of how L1 RT recognizes L1 and Alu RNA remains an open question. Since SINEs borrow retrotransposition enzymatic machinery from LINEs, the more general question is how RT recognizes LINE and SINE RNAs. The similarity between LINE and SINE elements was first shown for the turtle genome: it was found that elements from the LINE CR1 family share the same 3′-end with tRNA-derived SINEs from the same genome [20]. Later other examples were found and to date many LINE/SINE pairs, sharing the same 3′-end sequence have been identified [21, 22], including plants [23, 24], fish [25–27], insects [28], and mammals [29].

It was proposed to divide all LINEs in two groups: the stringent type and the relaxed type [22]. Transposons of the stringent type recognize their own 3′-end whereas transposons of the relaxed type do not involve any specific recognition of the 3′-end except for the poly-A tail. Because L1/Alu pairs do not share the same 3′-end sequence, the mammalian L1s were ascribed to the relaxed type. It has been suggested that the poly-A tail serves for RNA recognition and for the efficient L1-mediated retrotransposition [10, 18, 23, 30–32]. The direct monitoring of Alu retrotransposition in the human cells confirmed the requirement of a poly-A tail for Alu retrotransposition [18]. The phenomenon of poly-A tail expansion of new Alu insertions, presumably due to the slippage of the L1 ORF2 protein, was demonstrated in [33]. It was suggested that this effect may play an important role in maintaining activity of Alu elements [32]. It was experimentally confirmed that the poly-A tail is essential for L1 retrotransposition [14]. Recognition of poly-A tail by LINE retrotransposition machinery could explain formation of the processed pseudogenes [34], however many processed pseudogenes, lacking poly-A tails and derived from nearly all types of RNA, were found in various genomes [35, 36].

Co-localization of L1 RNA, ORF1 and ORF2 suggests that upon translation of both ORFs from L1 RNA they immediately form ribonucleoprotein particle (RNP) [37]. Additionally, it was discovered that L1 RT and RNA binding does take place and that it occurs at or near poly-A tail [37]. Branched molecules consisting of junctions between transposon 3′-end cDNA and the target DNA, as well as specific positioning of L1 RNA within ORF2 protein, were detected during initial stages of L1 retrotransposition in vitro [38]. Poly-A deletion did not have strong effect on retrotransposition while deletion of more substantial part reduced the number of transcripts [38].

The idea that secondary or tertiary RNA structure shared by L1 and Alu could be responsible for recognition and binding of ORF2, possibly along with a poly-A tail, was proposed by Boeke [30]. An important observation that retrotransposition may proceed even without a poly-A tail [35, 36] suggests presence of other important elements, playing a role in recognition of RNA by retrotransposition machinery.

For the stringent-type LINEs one of the important elements is a stem-loop structure located at the end of the 3′UTR [23–29]. Currently there is ample experimental evidence from different species that the 3′UTR stem-loop is essential for retrotransposition: for LINE SART1 in silkworm [28], for LINE UnaL2 in eel [39], for LINE ZfL2-1 and ZfL2-2 in zebrafish [40], for SINE SmaI in salmon [41] and for LINE R2 in insects [42]. The presence of a stem-loop at the 3′UTR of LINE and SINE, which share the same 3′-end, was found for a much broader range of species and LINE families, including L1 in algae [23] and monocot plants [24], Tad1 in fungi [43, 44], L2 in fish [25–27], and RTE in mammals [29]. These findings raise a question whether ORF2 functionality to recognize a stem-loop structure at the 3′-end of a stringent LINE/SINE pair was evolutionarily preserved for the relaxed type LINEs, and for L1s in particular.

An evolutionary study of LINEs, based on the RT domain of ORF2 protein, estimated that non-LTR retrotransposons are as old as eukaryotes, and revealed 11 distinct clades showing strict vertical descendance with no sign of horizontal transfer [45]. However, horizontal

transfer was reported for a minor fraction of LINE clades; these include transfer of jockey elements within Drosophila [46], Bov-B transfer from Squamata to the ancestor of Ruminantia [47, 48], insertion of additional C-terminal domain into ORF2 of insect R1 from plant viruses [49], and L1 transfer from humans to bacteria *Neisseria gonorrhoeae* [50]. Phylogenetic analysis made on the entire LINE ORF2 sequences, for which corresponding SINE partner is known [51], confirmed 11 clades reported by [45] and further enlarged the number of clades to 15 [51]. LINEs, which are known to share 3′-end sequences with SINEs, appeared to be enriched in L2, CR1, RTE and Ted1 clades. However, the dataset taken for analysis did not include plants and invertebrates, for which examples of SINE/LINE pairs were found too [22, 24, 52]. For plants, L1s, which share the same 3′-end sequence with SINE, were reported for green algae [23] and maize [24]. Phylogenetic analysis performed solely for the L1 clade and based on the entire ORF2 sequences, revealed monophyletic groups for green algae, land plants and vertebrates. Since land plants emerged from green algae, L1s of green algae show a strict mode of L1 recognition, and the strict-type L1s are observed in some plants, Ohshima proposed a model of parallel relaxation of stringent L1 RNA recognition in plants and mammals [51].

Evolutionary studies of L1 and Alu families in humans also presented evidence for the vertical evolution [53, 54]. Human L1 families were further subdivided into subfamilies, which were sequentially derived from a single lineage ending up in the currently active L1PA1 group of L1-Ta subfamily [53]. Analysis of Alu sequences led to identification of more than 200 families with 143 source elements [54].

Given that *(i)* evolution of LINEs in general and of L1s in particular showed mainly the vertical mode [21, 45, 53] with the stringent LINE/SINE pairs found among the different clades including L1; *(ii)* L1 RNA and ORF2 form stable RNP, and binding of the ORF2 to L1 RNA was reported to take place at or near poly-A tail [37, 38]; *(iii)* sequences, which lack poly-A tail, such as pseudogenes derived from different RNA genes, may undergo retrotransposition [35, 36], we suggest that L1 RNA recognition and binding with ORF2 could be evolutionarily preserved, though not at the sequence level but at the level of the RNA secondary structure. To test this hypothesis we investigated human L1 and Alu sequences for the presence of position-specific conserved stem-loop structures. We found highly conserved stem-loop position at the 3′UTR of L1 and at the 3′-end of Alu elements in human genome. Comparative analysis of this structure with other LINE 3′UTR stem-loop structures, which were experimentally shown to be essential for retrotransposition, revealed conservation of the structure without sequence homology. We found other conserved stem-loop positions at 5′UTR and at the end ORF2 proteins.

Results

L1

We performed an analysis of human L1 and Alu transposon sequences for the presence of position-specific stem-loop structures. Analyses were done for sets of presently active L1 transposons taken from [3], a set of 6622 L1s, divided into 27 subfamilies, as reported in [53] (the coordinates of all L1s used in this study are given in Additional file 1), and a set of 401 242 Alus, divided into 213 subfamilies as reported in [54] (coordinates of all Alus used in this study are given in Additional file 2). All transposon sequences were annotated with stem-loop structures, and this annotation was used for the construction of stem-loop coverage profiles (see Methods). High values in the stem-loop coverage profiles correspond to the conserved stem-loop positions.

The stem-loop coverage profiles for different sets of L1 sequences are shown in Fig. 1. Figure 1a depicts conservation profiles for 6 most active L1 transposons taken from [3]; the profiles for the representative L1 family clades – L1PA, L1PB and L1MA – are presented in Fig. 1b-e. The conserved stem-loop positions of the most recent L1PA1 subfamily coincide with the stem-loop positions of 6 most active transposons. The level of position conservation gradually decreases with the age of L1 subfamily due to insertions, deletions and 5′UTR truncations, which affect transposon length, and also due to mutations affecting the stem-loop structure. The stem-loop conservation profiles for all 27 L1 subfamilies from [53] are provided in Additional file 3 in the order following the phylogenetic tree depicted in Fig. 2 in [53]. This trend of relaxation of position conservation is clearly seen in the direction from younger to older families and is in agreement with L1 vertical evolution. The profiles of 6 most active L1s and of the enlarged set of 33 active L1s, also reported in [3] are almost identical (Additional file 4).

Reconstructed stem-loop conservation profiles revealed presence of conserved positions along the entire transposon length at 5′UTR, ORF2 and 3′UTR. According to the profiles of currently active L1s and L1PA1 subfamily, we distinguished three characteristic regions at 5′UTR (5′UTR-1, 5′UTR-2 and 5′UTR-3), two characteristic positions at the end of ORF2 (ORF2-1 and ORF2-2) and, importantly, a conserved position at 3′ UTR. All these positions are discussed below.

Although phylogenetic analysis of L1 ORFs sequences shows evolution of a strictly vertical type, this is not the case for non-coding parts of L1, 3′UTR and 5′UTR, which are not conserved at the sequence level. L1 changed 5′UTR several times in the course of evolution [53]. We analyzed 5′UTR regions of groups of L1 subfamilies

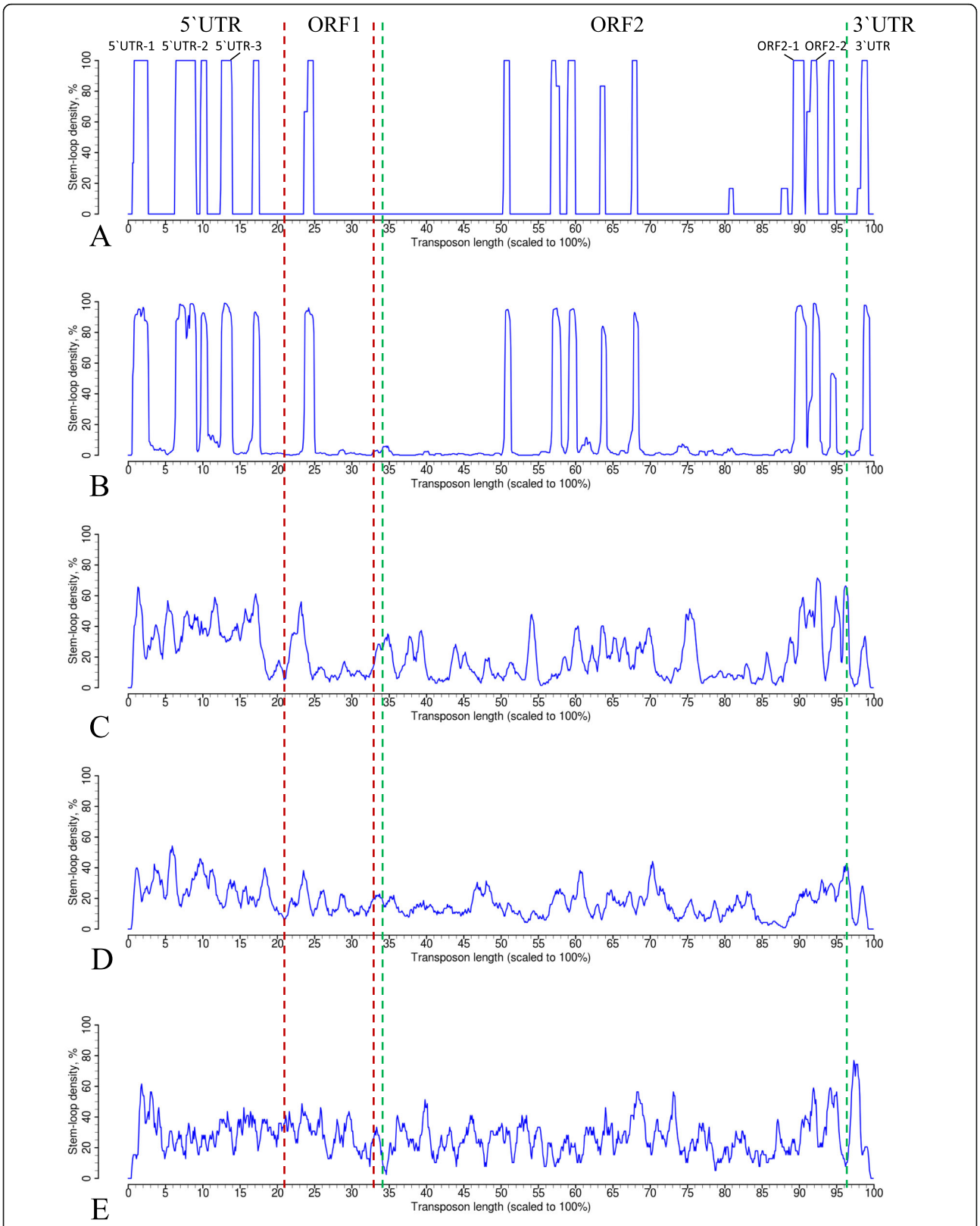

Fig. 1 Stem-loop coverage profiles. **a** 6 hottest L1 transposons reported as active in [2]; **b** L1PA1, the most active L1 subfamily from [46]; **c** L1PA8, middle-aged subfamily from L1PA clade [46]; **d** L1PB1, subfamily from L1PB clade [46]; **e** L1MA1, subfamily from L1MA clade [46]

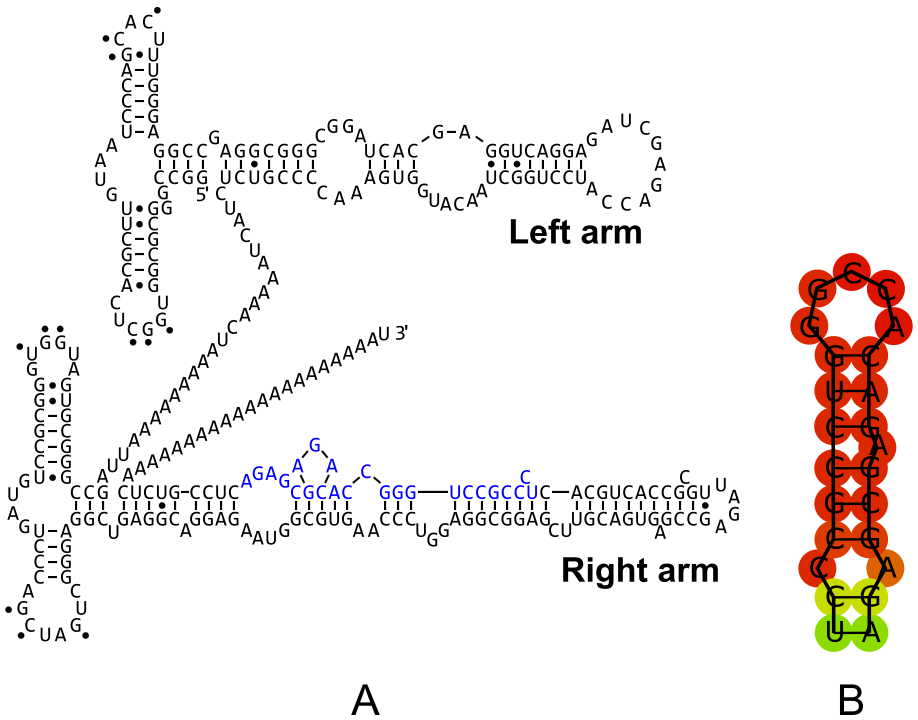

Fig. 2 Alu and 3′-end stem-loop secondary structure for AluYa5 active element. **a** Alu secondary structure. Alu is composed from two monomers of 7SL RNA. Each monomer has two short stem-loops connected with U-turn at the 5′-domain and a long stem-loop with bulges at the 3′-domain. 3′-end sequence that can form stem-loop is highlighted in blue. **b** Alu 3′-end stem-loop structure from the sequence highlighted in *blue* in the section (A)

having one type of 5′UTR as it was proposed in [53]. The stem-loop conservation profiles for 5′UTR regions for L1 subfamilies are presented in Additional file 5. The youngest group includes L1PA subfamilies from 1 to 8, and the profile reflects four conserved positions along 5′ UTR region. As expected, significantly less conservation is observed for the older groups.

The coordinates of stem-loops along the entire transposon length for the consensus sequence of each L1 subfamily are given in Additional file 6. We also checked how the mutations, which lead to the corrupted ORFs affect stem-loop structures. The stem-loop profiles of L1s with intact and corrupted ORFs are given in Additional file 7. At least half of the mutations affecting ORFs also affect stem-loop structures in the ORF1 and ORF2 regions. However 3′UTR region remained the most conserved in terms of stem-loop structure conservation compared to 5′UTR region.

Alu

The RNA secondary structure of Alu monomers is thought to be the same as that of 7SL RNA since Alu is composed of two 7SL Alu-domains [17]. Each monomer has two short stem loops connected with U-turn at 5′-domain and a long stem-loop with bulges at 3′-domain [17] (Fig. 2 depicts active AluYa5 element). Mutagenic experiments with Alu monomer binding affinity to SRP9/14 showed that retrotransposition is affected more by the mutations in the binding region of the left monomer, while the mutations in the binding region of the right monomer do not have a strong effect on retrotransposition. This finding suggests different roles two monomers may play in retrotransposition. Most likely, while the left Alu monomer is bound to SRP9/14, the right monomer participates in the recruitment of LINE RT, of which process the stem-loop recognition could be an important step. 7SL RNA structure contains three stem-loops, two in 5′-domain and one in 3′-domain, and each potentially could be used in binding with LINE RT. However the reverse transcription starts precisely from the 3′-end of Alu, and that is why the structures, close to 3′-end could be essential for binding with the RT.

Here we performed an analysis of 401 242 Alu elements, divided into 213 subfamilies, as reported in [54], for the presence of a 3′-end stem-loop structure. For all the elements from all families, AluJ, AluS and AluY, we detected a potential stem-loop structure, located at the very end of 3′-end region, several nucleotides before the poly-A tail. However, if the right Alu monomer accepts the known 7SL RNA structure, the predicted 3′-end stem-loop is hidden and its entire palindromic sequence is a part of the stem of the right arm (the sequence

highlighted in blue in Fig. 2b). The structure of this stem-loop is given in Fig. 2b and it has a 5 bp central loop together with an internal symmetrical loop (1–1) located at a distance of 6 bp from the central loop.

We took consensus sequences for each of 213 Alu subfamilies, constructed multiple alignment and built a sequence logo profile (see Fig. 3). The level of sequence conservation for the left Alu monomer is higher than that for the right. The position of 3′-end stem-loop structure, which reveals some degree of variation, is highlighted in red. The small region of 7 bp to the left of the stem-loop with a low conservation is because of the insertion into two highly active Alu elements, AluYB8 and AluYB9 [55]. The coordinates of 3′-end stem-loop structures for the consensus sequences of each subfamily is provided in Additional file 8. We analyzed structural features of the predicted stem-loops depending on the subfamilies and found the following trend. All of the predicted stem-loops fall into two major classes – those with a bulge and those without a bulge. Almost all stem-loops from the ancient families, AluJ, have a 3′-end stem-loop without a bulge, while AluY contains almost all 3′-end stem-loops with a bulge, similar to those presented in Fig. 2. Proportions of stem-loops with and without a bulge in the middle-aged AluS family are almost equal. A possible role of a bulge in stem-loop structures in general is discussed below.

LINE 3′ UTR stem-loop

3′ UTR region occupies ~200-245 bp at the end of the L1 transposon immediately after ORF2. Stem-loop density profiles (Fig. 1) revealed that active transposons and the youngest L1PA1 family have a distinct peak in the 3′UTR region. For different sets of active L1 elements we extracted sequences corresponding to 3′UTR stem-loops and reconstructed their secondary structures. For two sets of experimentally confirmed active transposons from [3] the corresponding stem-loop is located within the last 50 bp of L1. 3′UTR stem-loop secondary structure is presented in Fig. 4a.

Then we extracted 3′UTR stem-loop structures for different classes of LINEs L1 3′UTR stem-loop from 6 hottest human L1 transposons reported as active in [3] (Fig. 4a); L2 3′UTR stem-loop from eel [26] (Fig. 4b); two L2 3′UTR stem-loops from zebrafish (ZfL2-1 and ZfL2-2) [40] (Fig. 4c-d); R1 3′UTR stem-loop from silkworm experimentally reported as recognized by LINE-encoded RT [28] (Fig. 4e); L1 3′UTR stem-loop from rat [56] (Fig. 4f); L1 3′UTR stem-loop from maize [45] (Fig. 4g); and R1 3′UTR stem-loop from mosquito [45] (Fig. 4h). Four of them: L2 from eel (UnaL2) [26], L2 from zebrafish (ZfL2-1 and ZfL2-2) [40], and R1 from silkworm (SART1) [28] were experimentally reported as participating in RT binding. The characteristic feature of the reported stem-loops is an asymmetrical internal loop, or a bulge, located at a distance of 4–6 bp from the central loop (Fig. 4a-h).

We examined the structure of 3′UTR stem-loop located at the very end of 6 L1 human hot transposons and found that it has a structure with a bulge most similar to that of a zebrafish. By structural similarity we mean the length of the loop and position of a bulge with

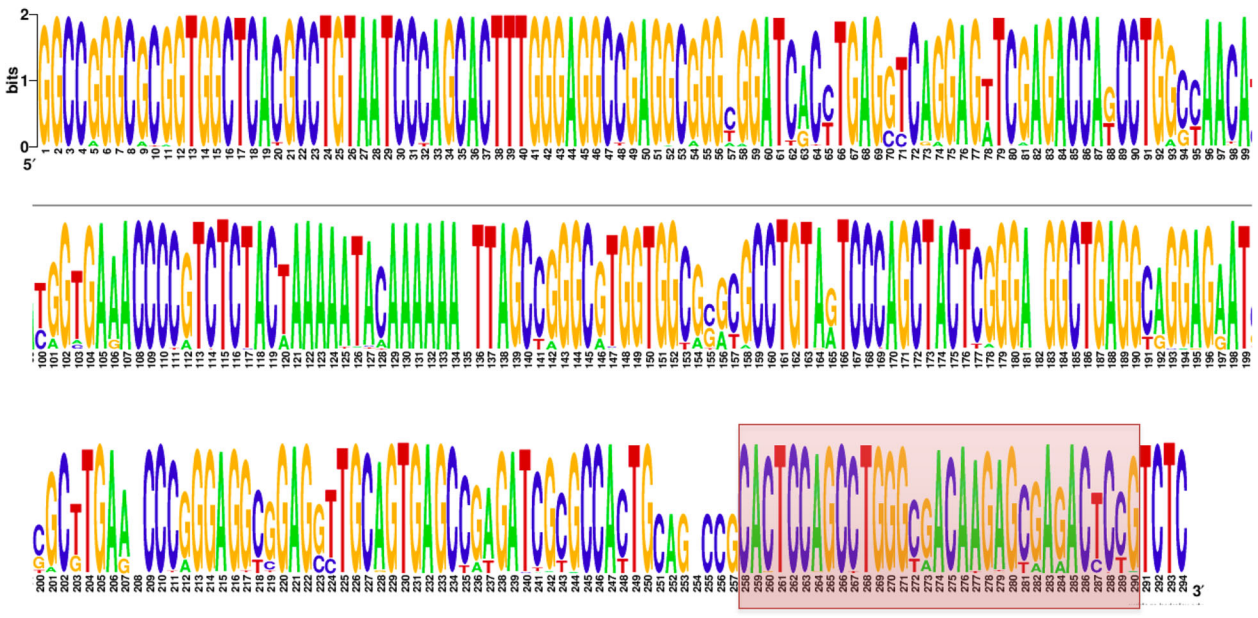

Fig. 3 Alu sequence logo profile constructed from consensus sequences of 213 Alu families. Position of 3′-end stem-loop structure is highlighted in *red*

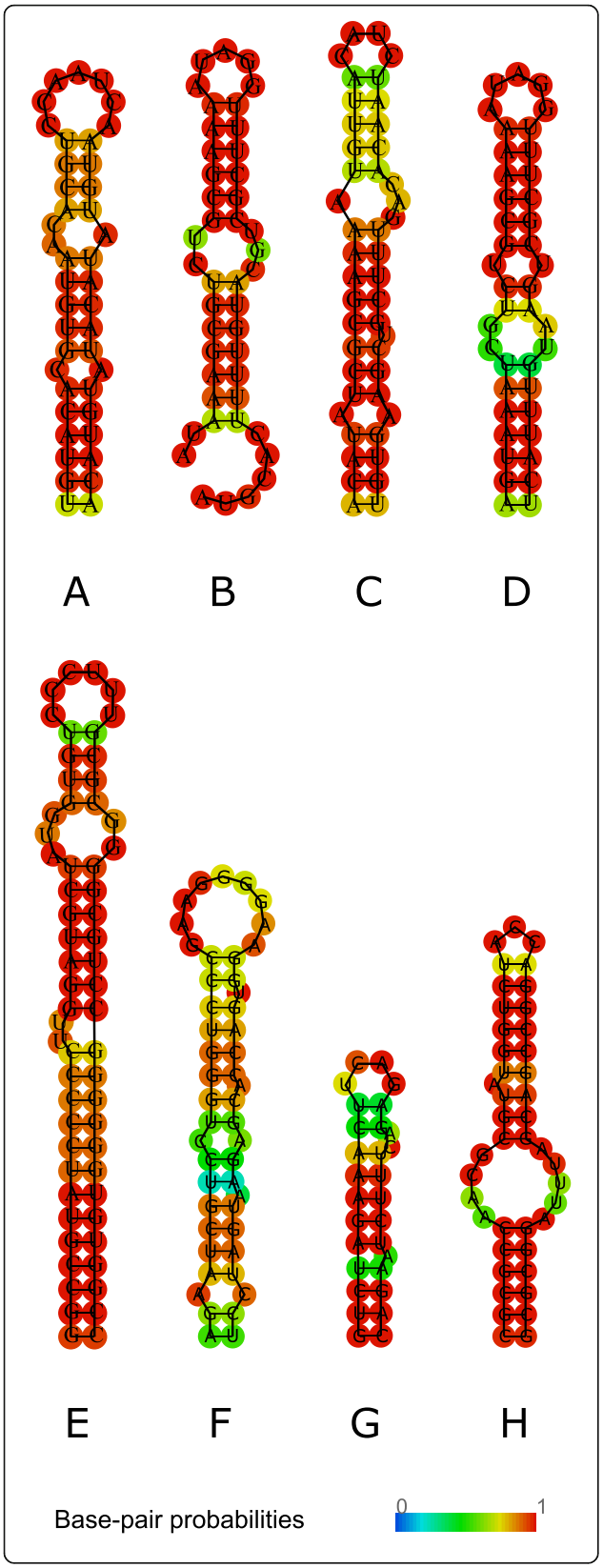

Fig. 4 RNA secondary structures for 3′UTR stem-loops from different species. **a** L1 3′UTR stem-loop from human genome taken from 6 hottest L1 transposons reported as active in [2]. **b** L2 3′UTR stem-loop from eel experimentally reported as recognized by LINE-encoded RT [26]. **c**-**d** two L2 3′UTR stem-loops from zebrafish (ZfL2-1 and ZfL2-2) experimentally reported as recognized by LINE-encoded RT [40]. **e** R1 3′UTR stem-loop from silkworm experimentally reported as recognized by LINE-encoded RT [28]. **f** L1 3′UTR stem-loop from rat [49]. **g** L1 3′UTR stem-loop from maize (Malik, Burke et al. [45]). **h** R1 3′UTR stem-loop from mosquito (Malik, Burke et al. [45])

respect to the loop. We also reconstructed secondary structures for 3′UTR stem-loops from the set of 33 active L1s and 6622 L1s. The characteristic stem-loop with a bulge was found in all 33 active L1s and in more than 50% of 6622 L1s. In other cases, the structure represented a long stem-loop up to 11–12 bp with one unpaired base at one side of the stem at the position of the bulge, and in some cases with 5 up 10 asymmetrically unpaired bases in the bulge.

To assess the distribution of the bulge size and position in L1 and Alu 3′-end stem-loops we constructed pairing/unpairing profiles (see Methods) where non-zero values correspond to the unpaired bases (Fig. 5). The profiles for L1 (Fig. 5a-b) show the presence of two internal loops, when the first internal loop occupies region from 4–5 bp if counting from the central loop. For Alu we compared two structures – one that is formed by the 3′-end, and the second one within the last 50 bp of the 3′-end. The profile of the 3′-end stem-loop (Fig. 5c-d) revealed the presence of three unpaired regions located at 3 bp, 6–7 bp and 12 bp from the central loop. For the stem-loop corresponding to the end of the right arm single unpaired nucleotides are located at 4 bp and the bulge positions are 13–14 bp (Fig. 5e-f). The profiles showed that the L1 3′-end stem-loop profile is more similar to the 3′-end Alu stem-loop rather than to the stem-loop at the end of the right arm.

LINE ORF2 stem-loop binding region

The currently accepted model of LINE RNA recognition by ORF2 is that there are two types of recognition – the stringent and the relaxed. In the stringent type RT recognizes its own 3′UTR tail, and in the relaxed type RT does not require any specific recognition except for the poly-A tail. Division into the stringent and the relaxed type came from the observation that some LINE/SINE pairs share the same 3′-end. For the stringent type, the experimental studies showed that a 3′UTR stem-loop is required for retrotransposition.

Evolutionary studies of LINEs for which corresponding SINE partners are known showed that in the phylogenetic tree, constructed from ORF2, the stringent- and the relaxed-type LINEs are intermixed, and the stringent-type LINEs are present in almost all branches including

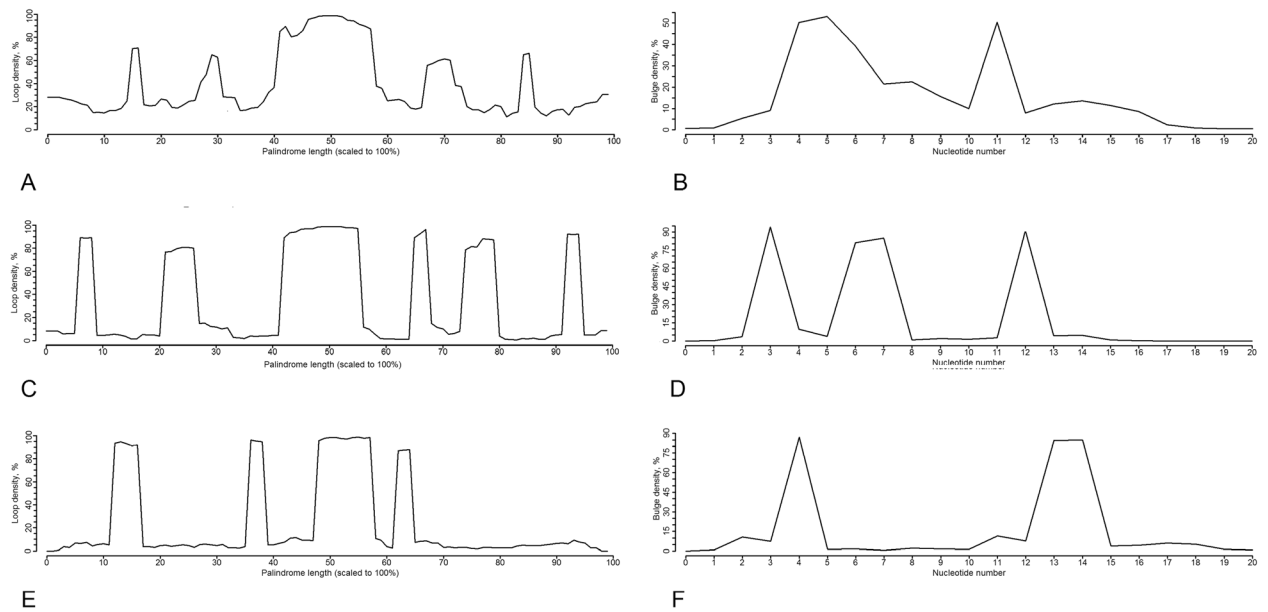

Fig. 5 Pairing-unpairing profiles for L1 3′UTR and 3′-end Alu stem-loops. The values of the resulted profile reflect how many stem-loop structures will have unpaired bases at a given position. Pairing-unpairing profiles for (**a**) L1 3′UTR stem-loop normalized to 100% length and (**b**) L1 3′UTR stem-loop, stem only; (**c**) stem-loop at the end of the Alu right arm normalized to 100% length and (**d**) stem-loop at the end of the Alu right arm, stem only; (**e**) Alu 3′-end stem-loop normalized to 100% length and (**f**) Alu 3′-end stem-loop, stem only

mammals, fishes, insects and plants (see phylogenetic trees in [21, 51]).

It was shown that for zebrafish LINEs ZfL2-1 and ZfL2-2, belonging to the stringent type L2 clade, the region of the ORF2 that binds to the 3′-end lies between the endonuclease and reverse transcriptase domains [40]. Here we took ORF2 sequences from different LINE clades and investigated the region between EN and RT domains for the amino-acid conservation – approximately the region between 250 aa and 500 aa. The region of 250–500 aa of ORF2 alignment is presented in Fig. 6. Alignment is done for different types of LINE, it is not limited to L1 and includes L2, R1, R2, CR1, I, Jockey, and others. Although the level of conservation of the region between EN and RT domains is lower compared to the EN and RT domains (see full alignment in Additional file 9), it is still noticeable that the sequences are homologous. Therefore they could retain the function to recognize stem-loops. It is not excluded that 3′-end stem-loop recognition was conserved throughout evolution for all types of LINEs, both the stringent and the relaxed type.

5′ UTR and ORF2 stem-loops

The L1 stem-loop coverage profiles revealed three conserved stem-loop positions at 5′ UTR region, which we designated as 5′UTR-1, 5′UTR-2 and 5′UTR-3 (Fig. 1). The first conserved position at 5′UTR-1 corresponds to the very beginning of the transposon (~50-100 bp in the transposon coordinates). The stem-loop structure from the active transposons from this region is very stable, with a stem of 20 bp (Fig. 7a). It also has an internal asymmetrical loop (3–2) located at a distance of 10 bp from the central loop. The second conserved position (5′UTR-2) corresponds to the transposon region of ~420-600 bp. The structure located in this region is GC-rich and has a long stem of 17 bp. Its characteristic feature is a stretch of four C in the loop (Fig. 7b). The 5′UTR-3 region (around 600–840 bp) also contains GC-rich stem-loop structure with a stretch of 5 G-C pairs in the upper stem and 3 G-C pairs in the lower stem (Fig. 7c).

Two other stem-loop conserved positions are located at the end of ORF2-encoded sequence. The characteristic feature of these structures is that they contain short repeat sequences in the central loops. The region ORF2-1 (5397–5437 bp) contains two stem-loop structures. The first (5504–5536 bp) has poly (G) (4–5 G) in the central loop (Fig. 7d). The second ORF2-1 stem-loop contains TATA (or TATATA) repeat (Fig. 7e). The stem-loop structure from the ORF2-2 region (5540–5583 bp) is a long 16-bp structure with a bulge located at a distance of 6 bp from the central loop (Fig. 7f). The central loop contains CACA repeat.

Discussion

Retrotransposition is a multistage process that includes transcription, formation of ribonucleoprotein particles,

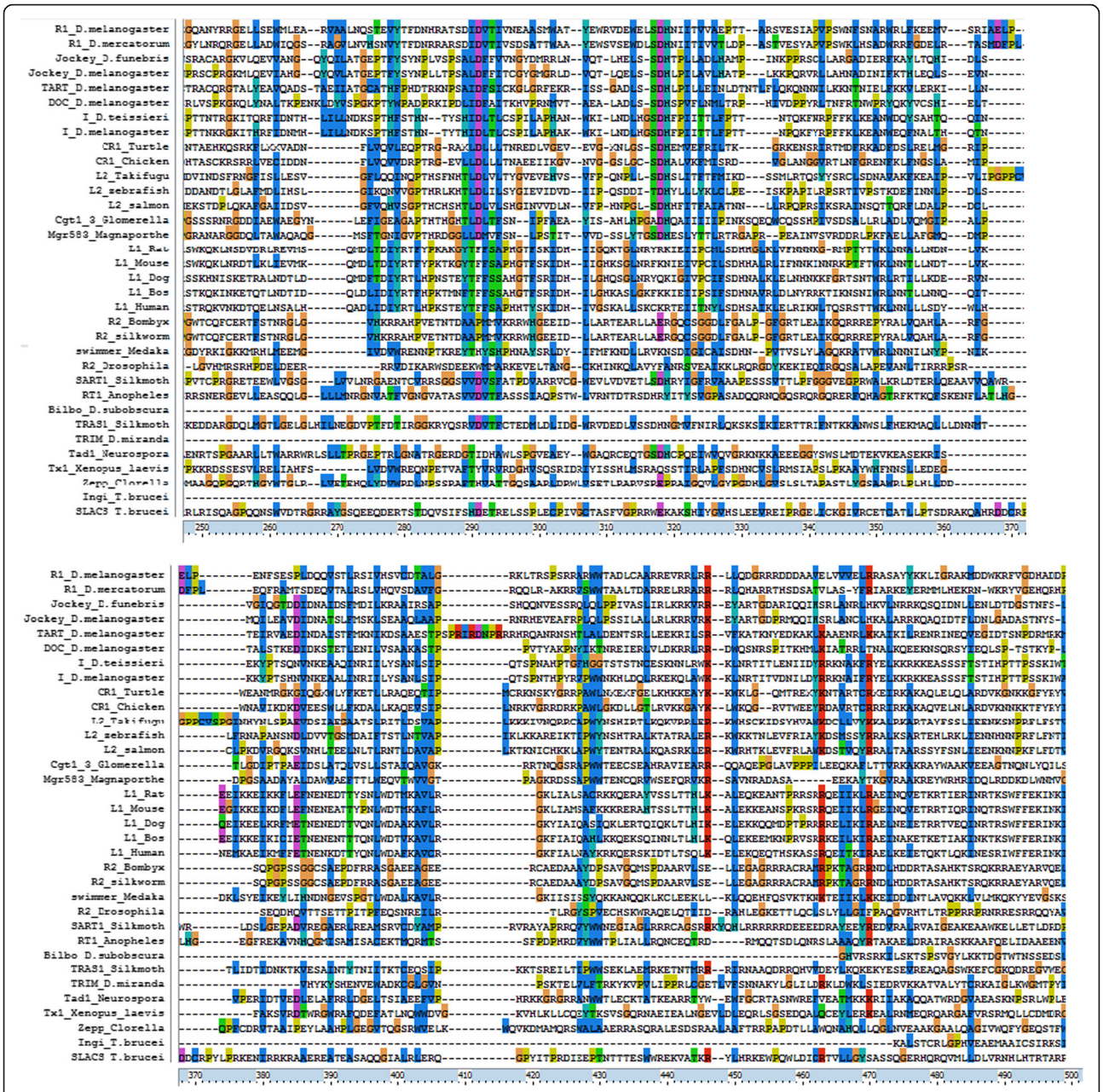

Fig. 6 Alignment of ORF2 region between EN and RT domains (250–500 aa)

translation, posttranslational modifications, transport back to the nucleus and integration in the DNA by the mechanism termed Target Primed Reverse Transcription (TPRT) [57]. Each stage involves transposon RNA interactions with different proteins including those encoded by the transposon itself, and often stem-loop structures participate in binding.

3′UTR. Stringent versus relaxed

The 3′UTR region was shown to play a crucial role in recognition of the stringent-type LINE RNA by ORF2 protein [40, 42, 58]. A stem-loop located at the very end of 3′UTR was confirmed to be an essential recognition motif for the stringent-type LINE-SINE pairs.

The LINE retrotransposon of insects, R2, requires 250 nucleotides at 3′UTR for recognition of its own RNA [42]. The R2 ORF2 protein from a silkworm *Bombyx mori* can identify 3′-end not only of its own RNA but also of R2 of *Drosophila melanogaster*. It is noteworthy that sequence similarity of this region in *Bombyx mori* and *Drosophila melanogaster* R2 is very low. Secondary structure models of 3′ UTR were predicted for both

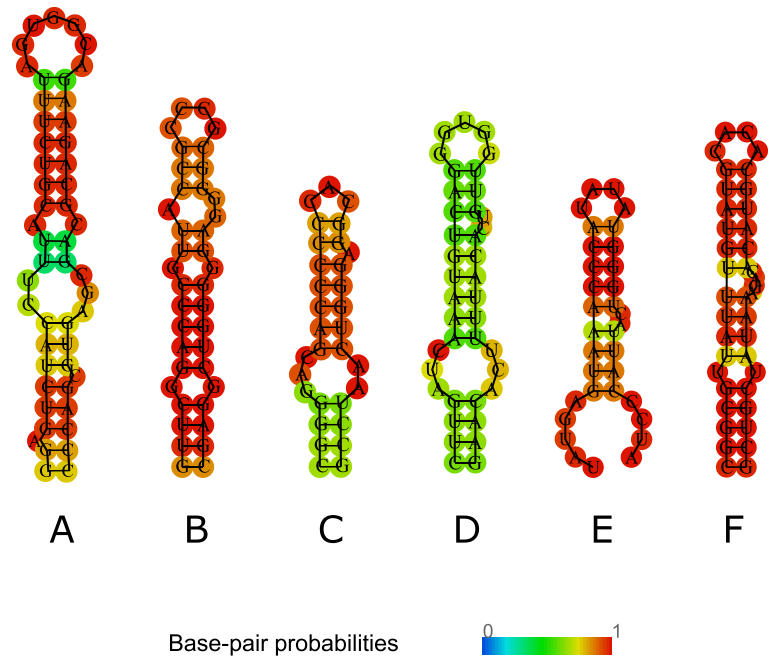

Fig. 7 RNA secondary structures of human L1 stem-loops from 5′UTR and ORF2 conserved positions. All the positions are highlighted in the L1 coverage profiles in Fig. 1. **a** 5′UTR-1 region, corresponds to ~50-100 bp. **b** 5′UTR-2 region, correspond to ~420-600 bp. **c** 5′UTR-3 region, correspond to ~600-840 bp. **d**-**e** ORF2-1 region, corresponds to 5397–5437 bp. **f** ORF2-2 region, corresponds to 5540–5583 bp

organisms and it appears that 3′UTR region contains several hairpins [59]. Later it was confirmed that for silkworm SART1 the central 3′UTR stem-loop is essential for retrotransposition, and that the transcription starts mostly from several telomeric repeat-like GGUU sequences just downstream of this stem-loop [28].

For eel genome it has been demonstrated that UnaL2 (LINE) and UnaSINE1 (SINE) share a similar 3′ tail, which is necessary for the successful UnaL2 transposition [26]. Moreover UnaL2 RT can recognize 3′ tail of UnaSINE1, thereby allowing its mobilization. The conserved 3′ tail consists of two parts: the stem-loop (a GGAUA loop) and the pentanucleotide repeat ([TGTAA]*n*). Both of them are necessary for the transposition of UnaL2. It has been suggested that 5′ part of the GGAUA loop is recognized by the RT [60]. The stem includes a small internal loop, which contains a single unpaired cytidine and U-U mismatch. Both are crucial for the successful transposition. This internal loop may contribute to the flexibility of the entire stem, which can be required for UnaL2 function [39].

In zebrafish ZfL2 ORF2 protein binds the hairpin located at 3′ tail via specific site between EN and RT domains [40]. Specifically, it was found that recognition can be bipartite, involving general recognition of the stem and specific recognition of the loop.

In salmon genome SINE SmaI element is derived from tRNA, and its 5′ region forms a tRNA-like cloverleaf structure while the 3′ domain forms an extended stem-loop structure, wherein the loop region is believed to be recognized by the LINE encoded RT [41]. It was also demonstrated that three other salmon SINEs, SlmI, HpaI and OS-SINE1 share the same 3′-end tail as LINE RSg-1 [27].

In turtle genome, 3′-end sequences of SINE family, designated as the tortoise PolIII/SINE, are also almost identical to 3′-end of LINEs from CR1 family [20]. This CR1-like LINE family is widespread in birds and in many other reptiles. Examples, in which 3′-ends of tRNA-derived SINEs are derived from 3′-ends of LINEs include transposons from turtles, fish, mammals and plants [22].

Homology between 3′-end of Vhc SINEs of bivalve mollusks and LINEs from CR1 clade of mollusk *Crassostrea gigas* was reported in [61]. Predictions of secondary structures of several Vhs SINEs from different mollusk species revealed the presence of 3′UTR stem-loops, however no experimental evidence is available to confirm the role of a 3′-end stem-loop in invertebrates. Another stem-loop from the central 40-bp subdomain of the V-domain was found to be conserved in SINEs in mollusks and arthropods, but its function remains unknown.

The experimentally confirmed evidence that a stem-loop structure at 3′UTR region, shared between LINE/SINE pairs, plays an important role in LINE ORF2 recognition raises a question about applicability of a similar type of recognition for the relaxed LINE-SINE pairs. A

phylogenetic analysis of ORF2 sequences showed the strict vertical evolution of this protein. It was shown that for zebrafish LINEs, the region recognizing 3′-end stem-loop lies between the EN and RT domains [40]. We showed that this region shows homology across different taxa and different LINE families. Importantly, the stringent LINE/SINE pairs are intermixed with the relaxed type across all branches of the phylogenetic tree starting from algae and protists, and are present in almost all divisions: in plants, insects, fish, reptiles and mammals.

We presented evidence that the recent human L1 and Alu families, to which the currently active retrotransposons belong, have a stem-loop structure at 3′-end of their sequences. We also identified presence of a 3′-end stem-loop structure in all (~400 000) analyzed human Alu sequences, however the formation of this structure *in vivo* is an open question. The tertiary structure of Alu is thought to be the same as that of two Alu domains of 7SL RNA bound by the RNA connector sequence. After Alu is transcribed it immediately forms ribonucleoprotein particle [62] together with the SRP9/14 heterodimer [55, 63], poly A-binding protein (PABP) [64] and perhaps some other proteins, which can bind RNA [19]. Crystal structures of SRP9/14 bound to the Alu domain revealed the exact positions of binding [65]. SRP9/14 binds strongly to the conserved core of 5′ domain, which forms a U-turn connecting two stem-loops (Fig. 1 in [65]). This part of 5′Alu-domain RNA is highly conserved in SRP RNAs from eubacteria to higher eukaryotes [66]. Mutagenesis experiments showed that the SRP binding regions of the two Alu monomers have different outcome, with the mutations in the right monomer having a minor effect on retrotransposition. This could mean that the right monomer can be used for RT recognition. While the left Alu-domain may be bound to SRP9/14, the protruding arm (positions 220–260) at 3′-domain of the right Alu monomer can potentially be used for RT recognition (Fig. 2). The structure of stem-loop at the end of the right arm (~220-260 bp) is structurally similar (in the sense that it also contains an internal bulge) to L1 3′UTR and zebrafish and eel LINE 3′UTR structures. For this structure to emerge, Alu right monomer has to acquire a conformation different from the accepted 7SL RNA. However, it is possible that the PBP bound to the polyA tail contributes to the unfolding of 3′domain, and this can lead to the formation of 3′-end stem-loop structure, that can be used for the recognition by the ORF2 protein.

Cis preference versus *trans* complementation

A number of authors sought to study the effect of preference of LINE machinery to various cellular RNA, such as LINE and SINE RNA and mRNA [30, 34, 37, 58]. It was shown that L1 proteins predominantly retranspose their own RNA – the effect named as *cis* preference [58]. In the same study it was demonstrated that L1 encoded proteins can retrotranspose cellular RNA, though at the much lower frequencies of 0.01-0.05% compared to the retrotransposition frequencies of the wild-type L1 RNAs. The mechanism through which retrotransposition of the non-LINE RNA occurs is termed *trans* complementation. It is believed that *cis* preference does not assume RNA recognition, and protein-RNA binding occurs cotranslationally due to ORF2 and L1 RNA proximity [58] while *trans* complementation requires a protein-RNA recognition both for Alus and pseudogenes. However mechanisms of the recognition remain unclear.

Even with *cis* preference it was demonstrated that L1 RNA, ORF1 and ORF2 form the ribonucleoprotein particle, and the binding of L1 RT takes place at or near the L1 RNA poly-A tail [37]. In the work of [34] *cis* preference was tested on the mutated L1 RNA transcript with stop codons in both ORFs and on the non-mutated L1 RNA, and no difference in the retroposition frequency was detected. This result supports the hypothesis of specific L1 RNA recognition by the encoded proteins. Nevertheless *cis* versus *trans* preference was confirmed in this study too. It should be emphasized that the noticeable effect of *cis* preference, where the proximity can facilitate binding, does not exclude the direct recognition of L1 RNA by the L1 encoded proteins. The presence of the secondary structure that participates in the binding of the RT with the retrotransposed RNA does not contradict to the effect of *cis* prefenece and *trans* complementation where both proximity and binding specificity can play an essential role.

Poly-A tail

Alu and L1 do not share the same 3′-end sequence except for the poly-A tail. It was believed that the poly-A tail is essential for the human Alu and L1 retrotransposition [30]. Direct monitoring of Alu retrotransposition in human cells revealed the functional importance of this element in Alu retroposition [18]. Poly-A tail is a part of Alu gene and not a result of polyadenylation, because Alus are transcribed by Pol III. Deletion of the poly-A tail from the DNA template almost abolished Alu transposition diminishing frequency by 1000 times [18]. An interesting observation of Alu tail expansion after the round of retroposition was reported in [33]. Expansion is thought to occur due to the slippage of L1 RT during the reverse transcription, and this effect could have provided evolutionary advantage to Alu elements to maintain their activity by counteracting to the

natural loss of A-tail [32]. Another finding of the same study was that the TPRT priming was not random with the priming positions being identified at least 25 bp from 5′-end of the poly-A tail [33]. This points to the constraints, which can be imposed by the bound proteins. Poly-A binding protein (PBP) can also bind Alu, and it was demonstrated that this protein is associated with SINE RNP [67].

The requirement of poly-A for the L1 retroposition was demonstrated in the work of [14]. Experiments were done with an engineered L1/MALAT RNA where L1 polyadenylation signal was replaced with 3′-end of the long non-coding RNA MALAT1, which can form a triple helical structure followed by a tRNA-like sequence. The triple helix can prevent RNA degradation in the absence of the poly-A tail [68]. L1/MALAT RNA lacking the poly-A tail is translated, but is not retrotransposed in *cis*. Addition of 16 or 40 poly-A sequence restored retrotransposition in *cis*. Also it was shown that poly-A tract is required for an association of the ORF2 and the retrotransposition-defective L1.

Whether the poly-A tail recognition is done involving the sequence directly or via the PBP remains an open question. Right after the translation of ORF1 and ORF2 both proteins associate with L1 and form RNP that goes back to the nucleus [69, 70]. It was found that PBP is associated with L1 ribonucleoprotein complex and is essential for retrotransposition [71]. Moreover, specific positioning of L1 RNA with ORF2 protein was observed [38]. Non-specific interaction of LINE RNA and ORF2 in human cells was found between the 180 aa carboxy-terminal segment (CTS) of L1 ORF2 protein and the human L1 RNA [72]. It was shown that a newly translated L1 ORF2 protein immediately binds with a high affinity to a native template with its C-terminal tail.

Some studies do not support the idea that poly-A tail is essential for retroposition. In the study of [38] the authors demonstrated *in vitro* that poly-A deletion did not have a strong effect on retrotransposition, while deletion of a more substantial 3′UTR region noticeably reduced the number of transcripts [38]. The abundance of retropseudogenes without poly-A tail also supports the idea that poly-A tail recognition can be bypassed in retrotransposition [35, 36]). These are tRNA-related tailless retropseudogenes, usually composed of 5′-part of the original tRNA or SINE founder RNA. Hundreds of thousands of tailless retropseudogenes derived from nearly all types of RNAs were discovered [35]. It was shown that L1-ribonucleoprotein particles are enriched in pseudogene transcripts [31]. These findings tell about the possibility for retrotransposition machinery to work without the poly-A tail, and for another essential recognition motif to be present in the retrotransposed sequences. This motif can be an RNA secondary structure. Here we hypothesize that an Alu element can mimic L1 with a structurally similar stem-loop though further experiments are needed to support this hypothesis.

Experimental cassettes studying retrotransposition

Assays to study retrotransposition are based on the idea that the indicators inserted in the cassettes will be expressed only upon a successful round of retrotransposition. Usually the indicators are inserted into 3′ UTR sequences. For example, *mneoI* cassette consists of the neo gene in the reverse orientation of the L1 transcription, inserted into 3′UTR, and the entire LINE-1 expression vector ends with SV40 polyadenylation signal [73]. This signal forms secondary structures, which were shown to be functionally important [74]. Specifically, a stem-loop structure, located downstream of the poly-A sequence correlates with the cleavage intensity [75]. Besides, a stem-loop structure was artificially added to the upstream region of *neo* gene in order to ensure binding of bacteriophage MS2 coat protein for the detection of L1 cellular localization by fluorescent in situ hybridization (FISH) [73].

Similarly, MALAT1, used in the experiments of [14] and disrupting the 3′UTR region, ends with a tRNA-like sequence, which is capable to form stem-loop structures. Thus, in all assays studying retrotransposition, 3′UTR region is disrupted by the insertions of various sequences, but all these sequences end with stem-loop structures.

Stem-loop structure and a bulge

Consensus structure of 3′-end transposon has an internal asymmetrical loop, or a bulge. Experiments with mutants from eel 3′-UTR stem-loop showed that the deletion of the bulge completely blocks the transposon activity [26]. Earlier it was shown for introns and viruses that the internal bulge in the stem of a stem-loop structure could be functionally important. For example, domain 5 (D5) of group II introns has a stem-loop structure with an internal bulge in the stem. It was shown that the loop, bulge and trinucleotide region in the lower stem are conserved features of splicing machinery of group II introns and also of spliceosome [76]. Specifically, the loop, the bulge and the triplet may bind essential metal ions to position functional groups participating in catalysis [76]. In HIV-1 the internal bulge of a stem-loop plays an important role in packaging through mechanisms, which are not fully understood [77]. Internal loops play an important role in discriminating between miRNA precursors and

other conserved hairpins [78]. Other examples where internal loops of stem-loop structures participate in binding include the murine IgM [79] and the yeast ribosomal protein L30 [80]. The study of thermodynamics for the reaction of a set of DNA hairpins containing internal loops showed that the size of an internal loop does matter and all targeting reactions proceed with negative changes in free energy, indicating that reactions proceed spontaneously [81].

For the detected 3′UTR stem-loop structure with a bulge we found that the position of a bulge is conserved in all active L1 transposons reported in [3] and conserved for more than 50% of the analyzed 6622 L1 elements from 27 families reported in [53]. Taking into account the experiments with eel transposons where deletion of a bulge completely abolished transposition, as well as other evidence of bulges playing an essential role, we hypothesize that the presence and location of a bulge in 3′UTR stem-loops plays an important role in RT recognition.

5′UTR

L1 transcription starts from its internal promoter. L1 5′UTR region contains internal promoter for two L1 ORFs, and it is not conserved at the sequence level. The length of the 5′UTR region is around 1000–1200 bp. Here we made analysis and present coordinates of the 6 active hottest transposons from [3]. First 100 bp were shown to possess promoter activity [5]. A binding site for the transcription factor YY1 was identified at positions ~3-26 bp with the core element sequence AAGATGGCC (~11-19 bp) [82]. The other binding sites (472–477 bp and 572–577 bp) for SRY family transcription factors were also identified [83]. Other transcription factors belonging to the family RUNX were shown to bind to 83–101 bp in 5′UTR region [84].

We found three conserved positions for stem-loops in 5′UTR regions of L1 active transposons belonging to L1P1 family. The first highly conserved 5′UTR-1 stem-loop (50–100 bp) is located within the 100 bp region reported as having the promoter activity, and it does not overlap with the binding sites reported for YY1 transcription factor, which are located in the first 50 bp area. The position of 5′UTR-2 stem-loop (423-461 bp) immediately precedes the SRY family binding site (472–477 bp), and there is a stem-loop (534-567 bp) located right before the second reported transcription factor binding site (572–577 bp).

Stem-loop structures in promoter regions with an important functional role were reported mostly for viruses [85, 86]. Perfect palindromes with the stem length of 5 bp were found in TATA-less promoters of ~ 5% of human genes [87]. Comprehensive analysis of five potential promoters of the HNRNPK gene showed that the one containing a palindrome [88] showed the highest activity. Further experimental efforts are required to study the role of stem-loop structures located at 5′UTR regions of the transposable elements.

ORF2

Stem-loop structures found at the end of ORF2 gene have characteristic dinucleotide repeats TATA and CACA, and also poly (G) sequence in the central loop, suggesting a possible role these stem-loops may play in the recognition by specific proteins. The role of the CACA-repeat as a regulator of the mammalian alternative splicing was revealed in [89]. Dinucleotide repeat motifs were found to be enriched in enhancers in Drosophila [90]. It is possible that position-specific stem-loops at the end of ORF2 protein play a role in the formation of ribonucleoprotein complexes, which direct the transport of L1 RNA into specific cell locations. Little is known about mRNA structures *in vivo*. The recent study of mRNA structures revealed abundance of intra-molecular double-stranded RNA [91]. Moreover, depletion of the coding regions and enrichment of 3′ UTR was observed. These results confirm the important role of RNA secondary structures in the post-transcriptional pathways of mRNAs, but further experiments are required to elucidate their function.

Conclusions

Here we presented an evidence for the presence of a highly conserved 3′UTR stem-loop structure in L1 and Alu transposons in human genome. We demonstrated that this 3′UTR stem-loop of L1 transposons is structurally similar to 3′UTR stem-loops of other LINEs from different species, which were experimentally reported as playing an essential role in retrotransposition [26, 28, 40], specifically RNA-binding region of ORF2 were determined in [26, 28, 40]. The region that binds to the 3′UTR stem-loop in zebrafish ZfL2-1 and ZfL2-2 transposons, shows homology across various LINEs from a wide range of species. The latter suggests that the functionality to recognize a stem-loop structure at 3′-end may have persevered through evolution for relaxed LINE/SINE pairs, including L1/Alu pair. Here we hypothesize that the binding of both L1 and Alu RNAs with L1 RT can be done via the structurally similar stem-loop structure at 3′-end of the transposon RNA. The other conserved stem-loop positions at 5′UTR and at the end of ORF2 suggest their possible functions in the protein-RNA interactions, but to date no experimental evidence has been reported.

Methods

Sets of L1 and Alu transposons

A set of 6622 L1 full-length elements divided into 27 subfamilies was taken from [53]. Consensus sequences for each subfamily were also taken from [53]. The sets of currently active L1 were composed from transposons experimentally reported as active in [3]. One set is composed from 33 reported active transposons, while the other set is made from a subset of 6 most active transposons, the activity of which is 10 times higher compared to other active elements [3].

A set of 401 242 Alus was taken from [54]. A division into 213 subfamilies and corresponding consensus sequences was based on the same study [54].

DNA Punctuation, a program that searches for stem-loop structures

A stem-loop structure is a palindromic structure that consists of a double-stranded stem, which is formed by completely or partially complimentary sequences, and a single-stranded loop. We define the following search parameters: minimum stem length (*minStemLen*), maximum stem length (*maxStemLen*), maximum loop length (*maxLoopLen*), and maximum number of mismatches in stems, which include gaps (*maxMismatch*). The task can be formulated as follows: given the sequence, find the longest complementary substrings in the range of *minStemLen, maxStemLen*, separated by a distance less than *maxLoopLen*, with no more gaps or mismatshes than *maxMismatch*.

The proposed algorithm is a dynamic programming algorithm based on the well-known Needleman–Wunsch sequence alignment algorithm [92]. However instead of looking for the alignment that produces a maximum score, we will look for the alignment with a minimum penalty score for gaps and mismatches. To work only with positive numbers we add nothing for matches and penalize with 1 for mismatches and gaps.

Thus, a matrix M, of size [(*minStemLen* +1) x (*minStemLen* +1)] is built as follows.

For $2 \le i \le minStemLen + 1$, $2 \le j \le minStemLen + 1$:

$$M(i,j) = \min \begin{vmatrix} M(i\text{-}1, j\text{-}1) + S(i,j) (\text{mismatch or match}) \\ M(i, j\text{-}1) + w(\text{gap in sequence 1}) \\ M(i\text{-}1, j) + w(\text{gap in sequence 2} \end{vmatrix} \quad (1)$$

Where S (i,j) = 0 (match), S (i,j) = 1 (mismatch), w = 1 (gap penalty).

In the context of stems, a match corresponds to complementary nucleotides, while a mismatch refers to non-complimentary nucleotides. The first column and the first row are filled by substring coordinates. The values in the first column correspond to the coordinates of the first substring, and the values in the first row correspond to the coordinates of the second substring.

For example, for the two sequences TACG and AATGC, the initialized matrix is:

		A	**A**	**T**	**G**	**C**
	0	1	2	3	4	5
T	1					
A	2					
C	3					
G	4					

After initialization the matrix is gradually filled according to the scheme (1). For example, M [2, 2] = min (M [1, 1] + 0, M [1, 2] + 1,M [1, 2] + 1) = min (0,2,2) = 0.

		A	**C**	**T**	**G**	**C**
	0	1	2	3	4	5
T	1	**0** **(0+0,1+1,1+1)**	1 (1+1,0+1, 2+1)	2 (2+1,1+1,3+1)	3 (3+1,2+1,4+1)	4 (4+1,3+1,5+1)
A	2	1 (1+1,2+1,0+1)	1	1	2	3
C	3	2 (2+1,3+1,1+0)	2	1	2	3
G	4	3 (3+1,4+1,2+1)	3	2	2	2

After the matrix is filled the program searches for the cell with a maximum minimum of its coordinates – $\max_{ij}(\min(i,j))$, and whose value does not exceed *maxMismatch* – M(i,j) < = *maxMismatch*. If the corresponding coordinate is no less than *minStemLen*, the cell will be a starting point for the traceback processing. Let *maxMismatch* =2, and *minStemLen* =2, then the program searches for a maximum coordinate that has a value less or equal to 2. In the example this is the lowest right cell. The traceback processing is performed similar to the Needleman-Wunch algorithm with the only difference that the path is chosen according to minimum matrix values. The resulting stem structure is:

```
ACTGC
|-|||
T_ACG
```

To avoid cases where sequence pairing in stems starts with a gap or mismatch we add an additional requirement of M (1,1) = M (2,2).

Source code in C and the program *DNA Punctuation*, implementing the aforementioned algorithm, is available at www.dnapunctuation.org.

Stem-loop annotation and coverage profiles

Stem-loop structures were annotated with the program *DNA Punctuation* described above. We searched for structures with the stem length in the range of 15–50 bp, loop up to 10 bp, with 7 mismatches or gaps allowed. The stem-loop coverage profiles were constructed along the entire transposon length so that 1 is added if a base pair in the sequence is covered with a stem-loop and 0 otherwise. The total value for a given position was divided by the number of transposons, so that y-axis reflects a percentage of sequences having a base pair covered with the stem-loop. To adjust to a different length of the transposons we scaled the sequences to the normalized length of 100% with an average of 6 kb for L1 and 300 bp for Alu. In summary, the stem-loop coverage values reflect the percentage of analyzed sequences having sequence positions covered with stem-loop structures.

Structure analysis

Secondary structures for the selected stem-loop sequences were reconstructed and visualized with RNA fold software (Vienna RNA Package 2.1.9) [93].

To investigate position conservation for a bulge or an internal loop for a set of stem-loops we constructed pairing-unpairing profiles for a set of position-specific stem-loop sequences. If a base is unpaired then we add 1 to the profile or 0 otherwise. The values of the resulting profile reflect how many stem-loop structures will have unpaired bases at a given position. Two types of profiles were constructed: one with the length of a selected stem-loop normalized to 100%, and the other for a stem where only x-coordinates correspond to the exact position in a number of bases starting from the loop ($x = 0$) and towards the base of the stem.

Additional files

Additional file 1: Coordinates of 6622 L1s used in the study. (CSV 212 kb)

Additional file 2: Coordinates of 401242 Alus used in the study. (CSV 12945 kb)

Additional file 3: Stem-loop coverage profiles for 27 L1 subfamilies from (Khan, Smit et al. [53]) (PDF 2803 kb)

Additional file 4: Stem-loop profiles for active and highly conserved L1 transposons: (A) set of 6 hottest L1 transposons reported as active in (Brouha, Schustak et al. [3]); (B) set of 33 active L1 transposons reported as active in (Brouha, Schustak et al. [3]); (C) set of 6622 highly conserved L1 transposons (see Methods for selection criteria). (PDF 221 kb)

Additional file 5: Stem-loop profiles for 5′UTR regions of groups of L1 subfamilies having one type of 5′UTR as it was proposed in (Khan, Smit et al. [53]). (PDF 703 kb)

Additional file 6: Coordinates of stem-loops along the entire transposon length for the consensus sequence of each L1 subfamily. (CSV 472 kb)

Additional file 7: Stem-loop profiles for intact or non-intact ORFs: (A) set of L1 transposons with intact ORFs; (B) set of L1 transposons with frameshift mutations in ORFs. (PDF 217 kb)

Additional file 8: Coordinates of the 3′-end stem-loop for consensus sequences for 213 Alu subfamilies. (CSV 28 kb)

Additional file 9: Full alignment of ORF2 protein taken from different types of LINE, it is not limited to L1 and includes L2, R1, R2, CR1, I, Jockey, and others. (PNG 1346 kb)

Abbreviations

3′UTR: 3′ untranslated region; 5′UTR: 5′ untranslated region; EN: Endonuclease; FISH: Fluorescent in situ hybridization; LINE: Long interspersed element; ORF: Open reading frame; RNP: Ribonucleoprotein particle; RT: Reverse transcriptase; SINE: Short interspersed element; TPRT: Target primed reverse transcription

Acknowledgements

The calculations were partially performed at the Supercomputing Center of Lomonosov Moscow State University [94].

Funding

No funding for the research was received.

Availability of data and materials

Not applicable.

Authors' contributions

MP conceived the study and supervised the first author. DG performed all the calculations. MP and DG analyzed results and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Received: 20 July 2016 Accepted: 25 November 2016
Published online: 03 December 2016

References

1. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. Nat Rev Genet. 2009;10(10):691–703.
2. Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD. Molecular archeology of L1 insertions in the human genome. Genome Biol. 2002;3(10):research0052.
3. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian Jr HH. Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci U S A. 2003;100(9):5280–5.
4. Becker KG, Swergold GD, Ozato K, Thayer RE. Binding of the ubiquitous nuclear transcription factor YY1 to a cis regulatory sequence in the human LINE-1 transposable element. Hum Mol Genet. 1993;2(10):1697–702.
5. Swergold GD. Identification, characterization, and cell specificity of a human LINE-1 promoter. Mol Cell Biol. 1990;10(12):6718–29.
6. Holmes SE, Singer MF, Swergold GD. Studies on p40, the leucine zipper motif-containing protein encoded by the first open reading frame of an active human LINE-1 transposable element. J Biol Chem. 1992;267(28):19765–8.
7. Khazina E, Truffault V, Buttner R, Schmidt S, Coles M, Weichenrieder O. Trimeric structure and flexibility of the L1ORF1 protein in human L1 retrotransposition. Nat Struct Mol Biol. 2011;18(9):1006–14.
8. 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(3):731–6.

9. Januszyk K, Li PW, Villareal V, Branciforte D, Wu H, Xie Y, Feigon J, Loo JA, Martin SL, Clubb RT. Identification and solution structure of a highly conserved C-terminal domain within ORF1p required for retrotransposition of long interspersed nuclear element-1. J Biol Chem. 2007;282(34):24893–904.
10. Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian Jr HH. High frequency retrotransposition in cultured mammalian cells. Cell. 1996; 87(5):917–27.
11. Feng Q, Moran JV, Kazazian Jr HH, Boeke JD. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell. 1996;87(5):905–16.
12. Mathias SL, Scott AF, Kazazian Jr HH, Boeke JD, Gabriel A. Reverse transcriptase encoded by a human transposable element. Science. 1991;254(5039):1808–10.
13. Usdin K, Furano AV. The structure of the guanine-rich polypurine: polypyrimidine sequence at the right end of the rat L1 (LINE) element. J Biol Chem. 1989;264(26):15681–7.
14. Doucet AJ, Wilusz JE, Miyoshi T, Liu Y, Moran JV. A 3′ poly(A) tract is required for LINE-1 retrotransposition. Mol Cell. 2015;60(5):728–41.
15. Vassetzky NS, Kramerov DA. SINEBase: a database and tool for SINE analysis. Nucleic Acids Res. 2013;41(Database issue):D83–9.
16. Ullu E, Tschudi C. Alu sequences are processed 7SL RNA genes. Nature. 1984;312(5990):171–2.
17. Sinnett D, Richer C, Deragon JM, Labuda D. Alu RNA secondary structure consists of two independent 7 SL RNA-like folding units. J Biol Chem. 1991;266(14):8675–8.
18. Dewannieux M, Esnault C, Heidmann T. LINE-mediated retrotransposition of marked Alu sequences. Nat Genet. 2003;35(1):41–8.
19. Deininger P. Alu elements: know the SINEs. Genome Biol. 2011;12(12):236.
20. Ohshima K, Hamada M, Terai Y, Okada N. The 3′ ends of tRNA-derived short interspersed repetitive elements are derived from the 3′ ends of long interspersed repetitive elements. Mol Cell Biol. 1996;16(7):3756–64.
21. Ohshima K, Okada N. SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. Cytogenet Genome Res. 2005;110(1–4):475–90.
22. Okada N, Hamada M, Ogiwara I, Ohshima K. SINEs and LINEs share common 3′ sequences: a review. Gene. 1997;205(1–2):229–43.
23. Cognat V, Deragon JM, Vinogradova E, Salinas T, Remacle C, Marechal-Drouard L. On the evolution and expression of chlamydomonas reinhardtii nucleus-encoded transfer RNA genes. Genetics. 2008;179(1):113–23.
24. Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon JM, Westerman RP, Sanmiguel PJ, Bennetzen JL. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. PLoS Genet. 2009;5(11):e1000732.
25. Sugano T, Kajikawa M, Okada N. Isolation and characterization of retrotransposition-competent LINEs from zebrafish. Gene. 2006;365:74–82.
26. Kajikawa M, Okada N. LINEs mobilize SINEs in the eel through a shared 3′ sequence. Cell. 2002;111(3):433–44.
27. Matveev V, Nishihara H, Okada N. Novel SINE families from salmons validate Parahucho (Salmonidae) as a distinct genus and give evidence that SINEs can incorporate LINE-related 3′-tails of other SINEs. Mol Biol Evol. 2007;24(8):1656–66.
28. Osanai M, Takahashi H, Kojima KK, Hamada M, Fujiwara H. Essential motifs in the 3′ untranslated region required for retrotransposition and the precise start of reverse transcription in non-long-terminal-repeat retrotransposon SART1. Mol Cell Biol. 2004;24(18):7902–13.
29. Okada N, Hamada M. The 3′ ends of tRNA-derived SINEs originated from the 3′ ends of LINEs: a new example from the bovine genome. J Mol Evol. 1997;44 Suppl 1:S52–6.
30. Boeke JD. LINEs and Alus–the polyA connection. Nat Genet. 1997;16(1):6–7.
31. Mandal PK, Ewing AD, Hancks DC, Kazazian Jr HH. Enrichment of processed pseudogene transcripts in L1-ribonucleoprotein particles. Hum Mol Genet. 2013;22(18):3730–48.
32. Roy-Engel AM. A tale of an A-tail: the lifeline of a SINE. Mob Genet Elem. 2012;2(6):282–6.
33. Wagstaff BJ, Hedges DJ, Derbes RS, Campos Sanchez R, Chiaromonte F, Makova KD, Roy-Engel AM. Rescuing Alu: recovery of new inserts shows LINE-1 preserves Alu activity through A-tail expansion. PLoS Genet. 2012;8(8):e1002842.
34. Esnault C, Maestre J, Heidmann T. Human LINE retrotransposons generate processed pseudogenes. Nat Genet. 2000;24(4):363–7.
35. Noll A, Raabe CA, Churakov G, Brosius J, Schmitz J. Ancient traces of tailless retropseudogenes in therian genomes. Genome Biol Evol. 2015;7(3):889–900.
36. Schmitz J, Churakov G, Zischler H, Brosius J. A novel class of mammalian-specific tailless retropseudogenes. Genome Res. 2004;14(10A):1911–5.
37. Kulpa DA, Moran JV. Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nat Struct Mol Biol. 2006;13(7):655–60.
38. Cost GJ, Feng Q, Jacquier A, Boeke JD. Human L1 element target-primed reverse transcription in vitro. EMBO J. 2002;21(21):5899–910.
39. Nomura Y, Kajikawa M, Baba S, Nakazato S, Imai T, Sakamoto T, Okada N, Kawai G. Solution structure and functional importance of a conserved RNA hairpin of eel LINE UnaL2. Nucleic Acids Res. 2006;34(18):5184–93.
40. Hayashi Y, Kajikawa M, Matsumoto T, Okada N. Mechanism by which a LINE protein recognizes its 3′ tail RNA. Nucleic Acids Res. 2014;42(16):10605–17.
41. Kawagoe-Takaki H, Nameki N, Kajikawa M, Okada N. Probing the secondary structure of salmon SmaI SINE RNA. Gene. 2006;365:67–73.
42. Luan DD, Eickbush TH. RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element. Mol Cell Biol. 1995;15(7):3882–91.
43. Kachroo P, Leong SA, Chattoo BB. Mg-SINE: a short interspersed nuclear element from the rice blast fungus, Magnaporthe grisea. Proc Natl Acad Sci U S A. 1995;92(24):11125–9.
44. Meyn MA, Farrall L, Chumley FG, Valent B, Orbach MJ. LINEs and SINEs in Magnaporthe Grisea. In: 2nd international Rice Blast Conference Program Abstracts: 1998, 4–8 August Montpellier, France; 1998, 4–8 August, pp. 53.
45. Malik HS, Burke WD, Eickbush TH. The age and evolution of non-LTR retrotransposable elements. Mol Biol Evol. 1999;16(6):793–805.
46. Loreto EL, Carareto CM, Capy P. Revisiting horizontal transfer of transposable elements in drosophila. Heredity (Edinb). 2008;100(6):545–54.
47. Kordis D, Gubensek F. Unusual horizontal transfer of a long interspersed nuclear element between distant vertebrate classes. Proc Natl Acad Sci U S A. 1998;95(18):10704–9.
48. Gogolevsky KP, Vassetzky NS, Kramerov DA. Bov-B-mobilized SINEs in vertebrate genomes. Gene. 2008;407(1–2):75–85.
49. Lazareva E, Lezzhov A, Vassetzky N, Solovyev A, Morozov S. Acquisition of full-length viral helicase domains by insect retrotransposon-encoded polypeptides. Front Microbiol. 2015;6:1447.
50. Anderson MT, Seifert HS. Neisseria gonorrhoeae and humans perform an evolutionary LINE dance. Mob Genet Elem. 2011;1(1):85–7.
51. Ohshima K. Parallel relaxation of stringent RNA recognition in plant and mammalian L1 retrotransposons. Mol Biol Evol. 2012;29(11):3255–9.
52. Van Dellen K, Field J, Wang Z, Loftus B, Samuelson J. LINEs and SINE-like elements of the protist entamoeba histolytica. Gene. 2002;297(1–2):229–39.
53. Khan H, Smit A, Boissinot S. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res. 2006;16(1):78–87.
54. Price AL, Eskin E, Pevzner PA. Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. Genome Res. 2004;14(11):2245–52.
55. Bennett EA, Keller H, Mills RE, Schmidt S, Moran JV, Weichenrieder O, Devine SE. Active Alu retrotransposons in the human genome. Genome Res. 2008;18(12):1875–83.
56. 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(1–4):462–7.
57. Ostertag EM, Kazazian Jr HH. Biology of mammalian L1 retrotransposons. Annu Rev Genet. 2001;35:501–38.
58. Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, Boeke JD, Moran JV. Human L1 retrotransposition: cis preference versus trans complementation. Mol Cell Biol. 2001;21(4):1429–39.
59. Mathews DH, Banerjee AR, Luan DD, Eickbush TH, Turner DH. Secondary structure model of the RNA recognized by the reverse transcriptase from the R2 retrotransposable element. RNA. 1997;3(1):1–16.
60. Baba S, Kajikawa M, Okada N, Kawai G. Solution structure of an RNA stem-loop derived from the 3′ conserved region of eel LINE UnaL2. RNA. 2004;10(9):1380–7.
61. Matetovici I, Sajgo S, Ianc B, Ochis C, Bulzu P, Popescu O, Damert A. Mobile element evolution playing jigsaw - SINEs in gastropod and bivalve mollusks. Genome Biol Evol. 2016;8(1):253–70.
62. Bovia F, Fornallaz M, Leffers H, Strub K. The SRP9/14 subunit of the signal recognition particle (SRP) is present in more than 20-fold excess over SRP in primate cells and exists primarily free but also in complex with small cytoplasmic Alu RNAs. Mol Biol Cell. 1995;6(4):471–84.

63. Bovia F, Wolff N, Ryser S, Strub K. The SRP9/14 subunit of the human signal recognition particle binds to a variety of Alu-like RNAs and with higher affinity than its mouse homolog. Nucleic Acids Res. 1997;25(2):318–26.
64. Roy-Engel AM, Salem AH, Oyeniran OO, Deininger L, Hedges DJ, Kilroy GE, Batzer MA, Deininger PL. Active Alu element "A-tails": size does matter. Genome Res. 2002;12(9):1333–44.
65. Weichenrieder O, Wild K, Strub K, Cusack S. Structure and assembly of the Alu domain of the mammalian signal recognition particle. Nature. 2000;408(6809):167–73.
66. Strub K, Moss J, Walter P. Binding sites of the 9- and 14-kilodalton heterodimeric protein subunit of the signal recognition particle (SRP) are contained exclusively in the Alu domain of SRP RNA and contain a sequence motif that is conserved in evolution. Mol Cell Biol. 1991;11(8):3949–59.
67. West N, Roy-Engel AM, Imataka H, Sonenberg N, Deininger P. Shared protein components of SINE RNPs. J Mol Biol. 2002;321(3):423–32.
68. Wilusz JE, JnBaptiste CK, Lu LY, Kuhn CD, Joshua-Tor L, Sharp PA. A triple helix stabilizes the 3′ ends of long noncoding RNAs that lack poly(A) tails. Genes Dev. 2012;26(21):2392–407.
69. Doucet AJ, Hulme AE, Sahinovic E, Kulpa DA, Moldovan JB, Kopera HC, Athanikar JN, Hasnaoui M, Bucheton A, Moran JV et al. Characterization of LINE-1 ribonucleoprotein particles. PLoS Genet. 2010; 6(10). https://www.ncbi.nlm.nih.gov/pubmed/?term=20949108.
70. Kulpa DA, Moran JV. Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. Hum Mol Genet. 2005;14(21):3237–48.
71. Dai L, Taylor MS, O'Donnell KA, Boeke JD. Poly(A) binding protein C1 is essential for efficient L1 retrotransposition and affects L1 RNP formation. Mol Cell Biol. 2012;32(21):4323–36.
72. Piskareva O, Ernst C, Higgins N, Schmatchenko V. The carboxy-terminal segment of the human LINE-1 ORF2 protein is involved in RNA binding. FEBS Open Bio. 2013;3:433–7.
73. Richardson SR, Doucet AJ, Kopera HC, Moldovan JB, Garcia-Perez JL, Moran JV. The influence of LINE-1 and SINE retrotransposons on mammalian genomes. Microbiol Spectr. 2015;3(2):MDNA3-0061-2014.
74. Hans H, Alwine JC. Functionally significant secondary structure of the simian virus 40 late polyadenylation signal. Mol Cell Biol. 2000;20(8):2926–32.
75. Wu C, Alwine JC. Secondary structure as a functional feature in the downstream region of mammalian polyadenylation signals. Mol Cell Biol. 2004;24(7):2789–96.
76. Seetharaman M, Eldho NV, Padgett RA, Dayie KT. Structure of a self-splicing group II intron catalytic effector domain 5: parallels with spliceosomal U6 RNA. RNA. 2006;12(2):235–47.
77. Lawrence DC, Stover CC, Noznitsky J, Wu Z, Summers MF. Structure of the intact stem and bulge of HIV-1 Psi-RNA stem-loop SL1. J Mol Biol. 2003;326(2):529–42.
78. Ritchie W, Legendre M, Gautheret D. RNA stem-loops: to be or not to be cleaved by RNAse III. RNA. 2007;13(4):457–62.
79. Phillips C, Kyriakopoulou CB, Virtanen A. Identification of a stem-loop structure important for polyadenylation at the murine IgM secretory poly(A) site. Nucleic Acids Res. 1999;27(2):429–38.
80. White SA, Hoeger M, Schweppe JJ, Shillingford A, Shipilov V, Zarutskie J. Internal loop mutations in the ribosomal protein L30 binding site of the yeast L30 RNA transcript. RNA. 2004;10(3):369–77.
81. Prislan I, Lee HT, Lee C, Marky LA. The size of the internal loop in DNA hairpins influences their targeting with partially complementary strands. J Phys Chem B. 2015;119(1):96–104.
82. Minakami R, Kurose K, Etoh K, Furuhata Y, Hattori M, Sakaki Y. Identification of an internal cis-element essential for the human L1 transcription and a nuclear factor(s) binding to the element. Nucleic Acids Res. 1992;20(12):3139–45.
83. Tchenio T, Casella JF, Heidmann T. Members of the SRY family regulate the human LINE retrotransposons. Nucleic Acids Res. 2000;28(2):411–5.
84. Yang N, Zhang L, Zhang Y, Kazazian Jr HH. An important role for RUNX3 in human L1 transcription and retrotransposition. Nucleic Acids Res. 2003;31(16):4929–40.
85. Carpenter CD, Simon AE. Analysis of sequences and predicted structures required for viral satellite RNA accumulation by in vivo genetic selection. Nucleic Acids Res. 1998;26(10):2426–32.
86. Haasnoot PC, Brederode FT, Olsthoorn RC, Bol JF. A conserved hairpin structure in alfamovirus and bromovirus subgenomic promoters is required for efficient RNA synthesis in vitro. RNA. 2000;6(5):708–16.
87. Wyrwicz LS, Gaj P, Hoffmann M, Rychlewski L, Ostrowski J. A common cis-element in promoters of protein synthesis and cell cycle genes. Acta Biochim Pol. 2007;54(1):89–98.
88. Mikula M, Gaj P, Dzwonek K, Rubel T, Karczmarski J, Paziewska A, Dzwonek A, Bragoszewski P, Dadlez M, Ostrowski J. Comprehensive analysis of the palindromic motif TCTCGCGAGA: a regulatory element of the HNRNPK promoter. DNA Res. 2010;17(4):245–60.
89. Hui J, Hung LH, Heiner M, Schreiner S, Neumuller N, Reither G, Haas SA, Bindereif A. Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. EMBO J. 2005;24(11):1988–98.
90. Yanez-Cuna JO, Arnold CD, Stampfel G, Boryn LM, Gerlach D, Rath M, Stark A. Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. Genome Res. 2014;24(7):1147–56.
91. Sugimoto Y, Vigilante A, Darbo E, Zirra A, Militti C, D'Ambrogio A, Luscombe NM, Ule J. hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1. Nature. 2015;519(7544):491–4.
92. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48(3):443–53.
93. Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA package 2.0. Algorithms Mol Biol. 2011;6:26.
94. Voevodin VV, Zhumatiy SA, Sobolev SI, Antonov AS, Bryzgalov PA, Nikitenko DA, Stefanov KS, Voevodin VV. Practice of "lomonosov" supercomputer. Open Systems J. 2012;7:36–9.