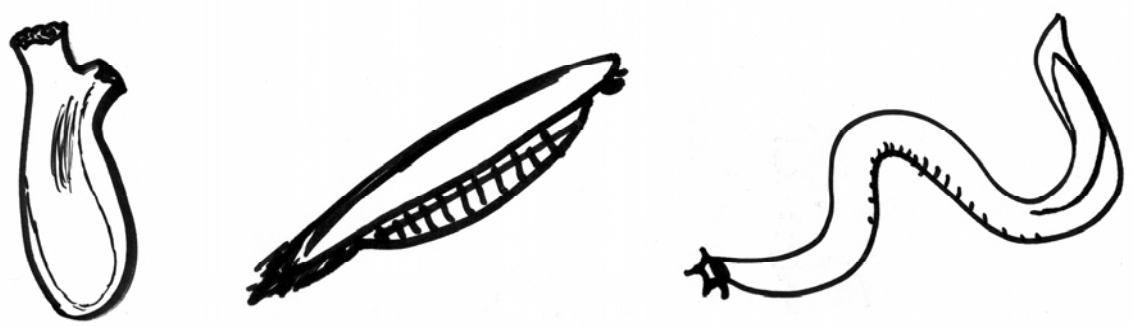


# Elements transposables de tipus non-LTR als ascidis, amfioxs i àgnats

Jon Permanyer i Ugartemendia

Programa de Doctorat del Departament de Genètica
Facultat de Biologia
Bienni 2002-2004

# Elements transposables de tipus non-LTR als ascidis, amfioxs i àgnats

Memòria presentada per
**Jon Permanyer i Ugartemendia**

Per optar al grau de
**Doctor**

Tesi doctoral sota la direcció del **Dr. Ricard Albalat Rodríguez** i
la **Dra. Roser Gonzàlez Duarte**
al Departament de Genètica de la Facultat de Biologia de la Universitat de Barcelona

Els directors L'autor

Dr. Ricard Albalat Dra. Roser Gonzàlez Jon Permanyer

Barcelona, Novembre 2006

# Publicacions

**Capítol 1: Retrotransposons non-LTR a *Ciona intestinalis***

Article 1: *The non-LTR retrotransposons in* Ciona intestinalis*: new insights into the evolution of chordate genomes*. **Genome Biol. 2003;4(11):R73**

Els urochordats són avui un bon model per entendre l'evolució dels organismes complexes ja que a nivell filogenètic s'els considera el grup germà dels vertebrats. La seqüenciació del genoma de l'ascidi *Ciona intestinalis*, un cordat no vertebrat, ha representat una fita en el emergent món de la genòmica comparada ja que ha posat a l'abast dels investigadors una informació bàsica per conèixer els elements (*tool kit*) dels genomes complexes i entendre un dels punts claus de l'evolució, l'origen dels vertebrats. Partint doncs de la base de dades d'aquest genoma ens varem proposar identificar la fracció corresponent als retrotransposons non-LTR i fer una anàlisi comparativa amb els elements descrits en d'altres espècies animals per establir-ne la participació en el llinatge dels cordats. Amb aquest objectiu s'han realitzat aproximacions *in silico* i experimentals que han permès llur caracterització. En el genoma de l'ascidi s'han descrit 5 seqüencies consens que presenten homologia a elements corresponents als clades I, L1, L2, LOA i R2. Els elements de l'ascidi conserven tant la pauta de lectura oberta corresponent a la transcriptasa inversa com l'estructura general corresponent a cada clade. A més a més se n'ha pogut determinar el número de còpies i l'estat de metilació de l'entorn genòmic d'aquests elements. La disponibilitat del genoma engalzat en *contigs* ha facilitat l'estudi de les seqüències flanquejants als elements descrits. Els resultats recollits en aquest treball evidencien una situació diferent a la observada als cordats vertebrats, clarament més propera a la d'altres protòstoms, ja que els retrotransposons non-LTR de l'ascidi són poc freqüents però molt variats i es troben en regions pobres en gens i preferentment localitzats en la fracció no metilada del genoma.

Research

Open Access

# The non-LTR retrotransposons in *Ciona intestinalis*: new insights into the evolution of chordate genomes

Jon Permanyer, Roser Gonzàlez-Duarte and Ricard Albalat

Address: Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Av. Diagonal 645, E-08028 Barcelona, Spain.

Correspondence: Ricard Albalat. E-mail: ralbalat@ub.edu

Published: 21 October 2003

*Genome* ***Biology*** 2003, **4:**R73

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/11/R73

Received: 5 August 2003
Revised: 17 September 2003
Accepted: 25 September 2003



## Abstract

**Background:** Non-long terminal repeat (non-LTR) retrotransposons have contributed to shaping the structure and function of genomes. *In silico* and experimental approaches have been used to identify the non-LTR elements of the urochordate *Ciona intestinalis*. Knowledge of the types and abundance of non-LTR elements in urochordates is a key step in understanding their contribution to the structure and function of vertebrate genomes.

**Results:** Consensus elements phylogenetically related to the I, LINE1, LINE2, LOA and R2 elements of the 14 eukaryotic non-LTR clades are described from *C. intestinalis*. The ascidian elements showed conservation of both the reverse transcriptase coding sequence and the overall structural organization seen in each clade. The apurinic/apyrimidinic endonuclease and nucleic-acid-binding domains encoded upstream of the reverse transcriptase, and the RNase H and the restriction enzyme-like endonuclease motifs encoded downstream of the reverse transcriptase were identified in the corresponding *Ciona* families.

**Conclusions:** The genome of *C. intestinalis* harbors representatives of at least five clades of non-LTR retrotransposons. The copy number per haploid genome of each element is low, less than 100, far below the values reported for vertebrate counterparts but within the range for protostomes. Genomic and sequence analysis shows that the ascidian non-LTR elements are unmethylated and flanked by genomic segments with a gene density lower than average for the genome. The analysis provides valuable data for understanding the evolution of early chordate genomes and enlarges the view on the distribution of the non-LTR retrotransposons in eukaryotes.

## Background

The ascidian *Ciona intestinalis* has joined the select group of fully sequenced genomes [1]. The draft sequence shows interesting features of an invertebrate chordate: a genome size of 153-159 megabases (Mb); base composition of 65% AT; 15,852 predicted transcripts; and a gene density of one per 7.5 kilobases (kb). *Ciona* genome organization lies between that of protostomes (most animals other than echinoderms and chordates) and vertebrates. The released sequence allows new approaches to study the structure of the still poorly characterized repetitive DNA fraction, which accounts for 30-35% of the urochordate genome [2]. Although rRNA and tRNA families have been described, the different classes of transposable elements were not surveyed. Indeed, current information about ascidian transposable elements is limited to only 1 Mb of genomic sequences [3]. These elements are,

however, invariably found in eukaryotes and most probably have contributed greatly to shaping the structure and function of vertebrate genomes [4].

Transposable elements are grouped into two major classes - class I and class II - depending on the mechanism of transposition [5,6]. Class I elements can be further classified into three categories: short interspersed nucleotide elements (SINEs); long terminal repeat (LTR) retrotransposons; and non-LTR retrotransposons (also termed LINE-like elements or retroposons). Although elements in the last category are among the most abundant, frequency estimates vary greatly depending on the species and the DNA segment considered, as most copies are 5'-truncated. Full-length non-LTR elements contain either one or two open reading frames (ORFs), all of them encode a reverse transcriptase, and some have additional motifs [7-10]. On the basis of the reverse transcriptase, non-LTR retrotransposons have been clustered into 14 different clades, the L1, L2, CR1, Rex1, RTE and R4 clades being the six major lineages present in vertebrates [11-15]. In contrast to vertebrates, our knowledge of LINE-like elements in other chordates is scanty: Cili-1 and Cili-2 [3] and BfCR1 [16] are the only non-LTR elements reported in non-vertebrate chordates. If, however, non-LTR clades originated before the divergence of the major animal phyla [11], urochordate and cephalochordate genomes should harbor representatives of these clades.

We have conducted an exhaustive search for non-LTR elements, initially on raw data and more recently on the draft genome, of the urochordate *C. intestinalis*. Phylogenetic analysis based on the reverse transcriptase domain showed that the ascidian elements grouped within five non-LTR clades. The structural features of the non-LTR elements, copy number, genome distribution and methylation status have been analyzed and inferences on the evolution of chordate genomes are presented.

## Results

### Non-LTR elements in the ascidian genome

Five consensus non-LTR retrotransposons, termed CiI, CiL1, CiL2, CiLOA and CiR2, were derived from five, five, six, five and five *C. intestinalis* scaffolds, respectively (Figures 1, 2 and see Additional data file 1). TBLASTX comparisons showed that the ascidian elements belonged to the I, LINE1, LINE2, LOA and R2 clades (E-values: $4e^{-69}$ with *Biomphalaria glabrata* (snail) BGR, $2e^{-89}$ with *Nycticebus coucang* (slow loris) L1, $2e^{-50}$ with *Danio rerio* CR1Dr2, $e^{-146}$ with *Aedes aegypti* Lian, and $e^{-106}$ with *Drosophila melanogaster* R2, respectively). CiL1.2 (Figure 2b), derived from five scaffolds, was another ascidian LINE1 element. It showed homology with the *Xenopus laevis* Tx1 retroelement (E-value: $2e^{-40}$) but was not further analyzed because it was significantly shorter than CiL1 (CiL1.2 only encompassed the reverse transcriptase region).

All the *Ciona* elements encoded the conserved reverse transcriptase with the distinctive structural hallmarks defined as block 0, 1, 2, 2a, 3, 4, 5, 6 and 7 [11] (Figure 2b). Although conservation of the thumb region (block 8 and 9) was weak, preservation of the ascidian sequences defining the CRE/R2/R4/L1/RTE subgroup was still found. The apurinic/apyrimidic endonuclease (APE) region was clearly identified in CiI, CiL1, CiL2 and CiLOA (Figure 2a) on the basis of the reported domains I to VII [17]. CiI and CiLOA also contained the RNaseH (RNH) domain at the carboxylic end (Figures 1, 2d). Concerning ORF1, partial sequences were assembled for CiI and CiLOA, but a CCHC motif (single-letter amino-acid code) in this region was only identified in the CiI element (Figure 2e). Finally, for CiR2, a restriction enzyme-like endonuclease (REL-endo) containing the CCHC and KPDI motifs [18] was found in the carboxy-terminal region, and a CCHH domain and a putative c-Myb DNA-binding motif were identified at the amino terminus (Figure 2c). Overall, the structure and organization of the ascidian non-LTR retrotransposons is consistent with those reported for each non-LTR clade.

### Phylogenetic relationships of *Ciona* retrotransposons

The reverse transcriptase domain of non-LTR elements was used to establish the phylogenetic relationships of the ascidian elements and the 14 reported non-LTR clades. In the neighbor-joining tree (Figure 3), all clades were supported with significant bootstrap values (• 70%), except clade I, with the lowest bootstrap value (67%) in agreement with previous analyses [11,12,14]. Therefore, ascidian sequences clustered within five distinct clades: I, L1, L2, LOA and R2 (bootstraps: 67%, 70%, 97%, 100% and 100%, respectively), as a result of which they were recorded as new members of such groups.

### Copy number and genomic features

Fragments of about 300 nucleotides of the reverse transcriptase domain of each ascidian element were PCR amplified, cloned, sequenced and used for copy-number estimations and methylation analyses. To quantify the copy number for each element, two independent experimental approaches - slot blot analysis and genomic library screenings (Figure 4a,b) - were combined with *in silico* scores on the number of *Ciona* scaffold-containing elements (Table 1). When the reverse transcriptase was considered, the data from the different approaches were consistent and mean values for each element were in the range 3-7, far below the copy numbers of the vertebrate counterparts (Table 2). CiR2 slot-blot analysis did not give a signal, in agreement with the low estimates obtained after *in silico* searches and library screenings. Indeed, full-length copies could not be assembled for any of the families after database searches. *In silico* estimates with sequences that also encompassed the 5' and 3' sequences of reverse transcriptase increased the numbers slightly: 9, 22, 24, 69 and 13 for CiI, CiL1, CiL2, CiLOA and CiR2, respectively (Table 1).

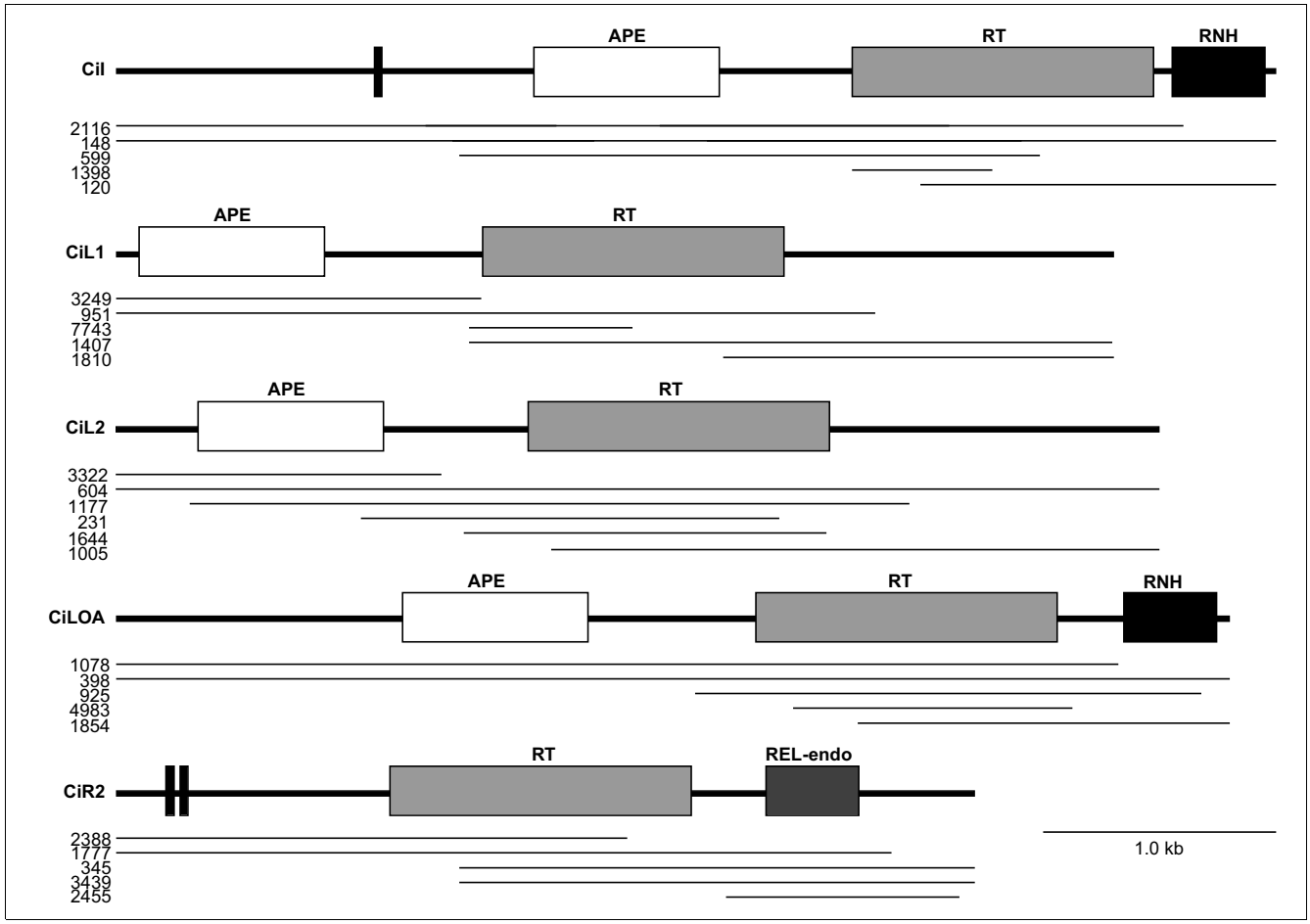



**Figure 1**
Schematic representation of the ascidian non-LTR retrotransposons. The conserved domains are depicted on the sequence derived from the *Ciona* scaffolds (thick line). Thin lines correspond to the physical segments covered by the scaffolds, which are numbered on the left. For clarity, indels are not shown. APE, apurinic/apyrimidinic endonuclease; REL-endo, restricted enzyme-like endonuclease; RNH, RNase H domain; RT, reverse transcriptase. Vertical bars indicate the location of cysteine-histidine motifs typical of nucleic acid-binding domains.

Gene density and GC content in the surrounding retrotransposon sequences was estimated from 26 10-kb regions flanking CiI, CiL1, CiL2 and CiLOA elements of 17 scaffolds. Overall, 16.5 genes were found in the 260 kb analyzed. Therefore, the average gene density (1 gene per 15.8 kb) was lower than that of the whole genome (1 gene per 7.5 kb) [1]. However, no differences were observed when the GC content of those segments (35.7%) was compared with the overall genomic value (35%). Concerning CiR2, our data confirmed the target specificity for rRNA genes associated with the REL-endo domain: 9 out of 13 CiR2s were indeed linked to rRNA sequences.

Finally, the methylation status of the genomic regions containing the elements was investigated by comparing the hybridization patterns of genomic DNA restricted with the methylation-sensitive enzyme *Hpa*II, and the methylation-insensitive isoschizomer *Msp*I. The identical *Hpa*II and *Msp*I patterns obtained for all the elements (except for CiR2, which gave no signal) supported the location of the ascidian elements in unmethylated genomic segments (Figure 4c).

## Discussion

### *Ciona* non-LTR retrotransposons

The analysis of non-LTR elements in the urochordate *Ciona* provides valuable data for understanding the evolution of early chordate genomes and enlarges the view of the distribution of the non-LTR clades in eukaryotes. The *Ciona* genome harbors: I, LOA and R2 elements, hitherto restricted to protostomes; L1 elements, formerly uncharacterized in invertebrates; and L2 elements, previously described in protostomes and vertebrates.

Clade I was the least supported branch of our analysis (bootstrap value, 67%). However, ascription of CiI to this clade was unambiguous as it shares with the other I elements the CCHC motif and the APE, reverse transcriptase and RNH domains

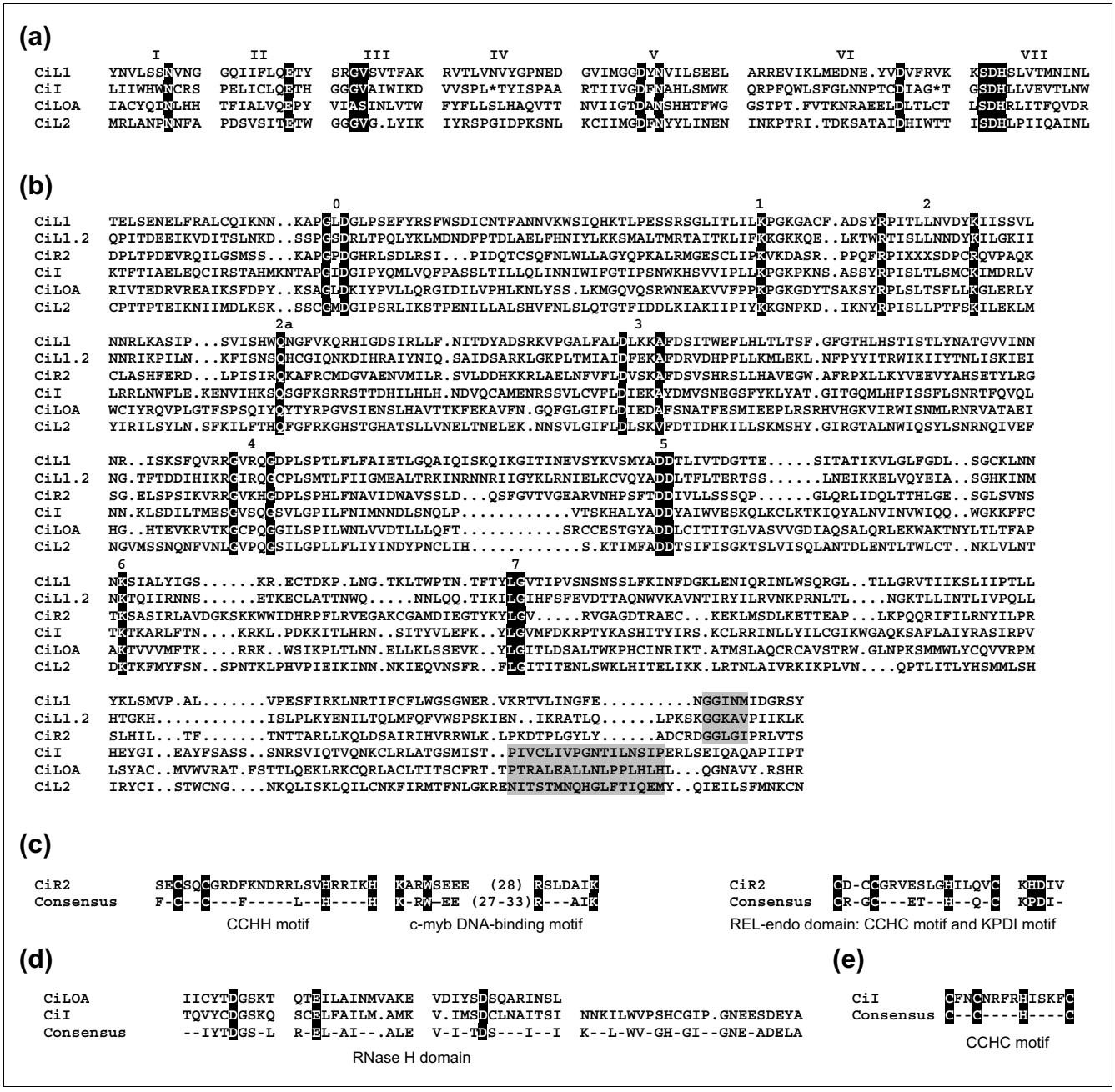



**Figure 2**

Multiple-sequence alignments of the *Ciona* non-LTR retrotransposons. **(a)** The APE region. Only blocks of highly similar residues are shown. The Roman numerals above the alignments correspond to those defined by Tu *et al.* [17]. Highly conserved residues, as convenient landmarks, are shaded. **(b)** Reverse transcriptase sequences. Numbers above the sequences and the black-shaded residues refer to the conserved peaks described by Malik *et al.* [11]. Gray-shaded residues correspond to the CRE/R2/R4/L1/RTE and Tad/R1/LOA/Jockey/CR1/I subgroups described in [11]. **(c)** CiR2 domains. The CCHH and c-Myb DNA-binding motifs are shown in the amino-terminal domain and the REL-endo domain with the CCHC and KPDI motifs in the carboxy-terminal domain [18]. **(d)** CiLOA and CiI RNH domains. Only the highly conserved regions of the RNase H domain of the blocks defined by Tu *et al.* [17] are depicted. The three amino-acid residues identified in the active site of *E. coli* RNase H are shaded. **(e)** CCHC motif of the putative CiI-ORF1 as defined by Fawcett *et al.* [32].

(Figure 1), and also because it clearly clustered with the I non-LTR retrotransposon BGR of the snail *Biomphalaria glabrata* (bootstrap 93%). In regard to the LOA clade, representatives in urochordates had previously been identified after BLASTN and BLASTX comparisons. The ascidian Cili-2 retrotransposon gives the closest match with the RNH domain of the mosquito Lian element [3]. We have now derived CiLOA, an element that encodes APE, reverse transcriptase and

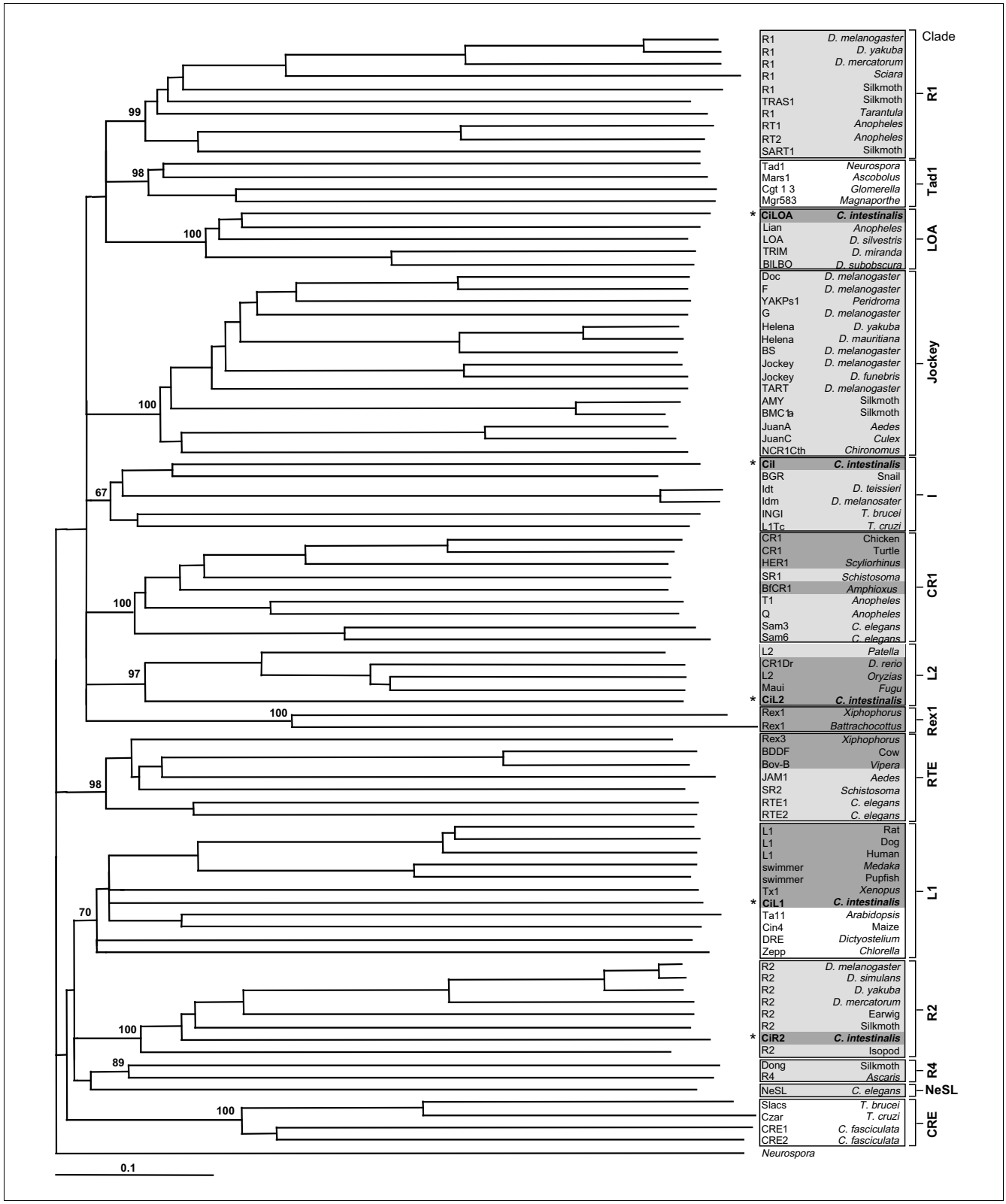



**Figure 3**
Phylogenetic tree of non-LTR elements based on the reverse transcriptase sequence. The elements identified in *C. intestinalis* are indicated with an asterisk. The number next to each node of the 14 clades indicates the bootstrap value as the percentage out of 1,000 replicates. The name of each non-LTR element and the species harboring it is listed to the right of the figure, shaded in light gray (protostomes) or dark gray (deuterostomes).

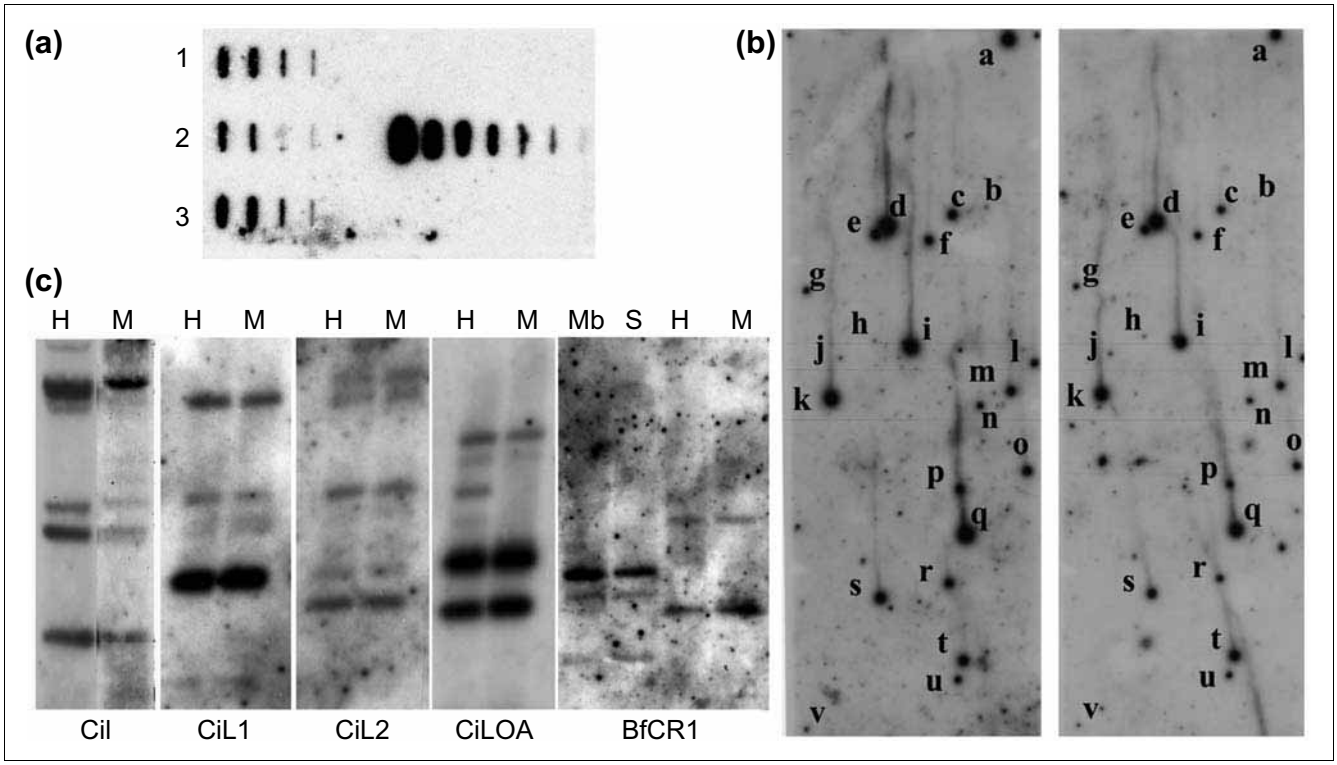



**Figure 4**

Slot-blot analysis, library screening and Southern blot of ascidian non-LTR elements. **(a)** On the left is shown a representative experiment of slot-blot analysis of CiLOA elements in three specimens with (from left to right) 500 ng, 250 ng, 50 ng and 25 ng of *Eco*RI-digested *C. intestinalis* genomic DNA. On the right is shown slot-blot analysis of serial dilutions of plasmid containing CiLOA which has been *Eco*RI-restricted and mixed with 1 μg mouse DNA. **(b)** Hybridization of a *C. intestinalis* genomic library screened with CiLOA. Positive signals have been depicted (from a to v) in the original (left) and its duplicate (right). **(c)** The first four panels show Southern analyses of 10 μg *C. intestinalis* genomic DNA digested with *Hpa*II (H) or *Msp*I (M) and probed with the non-LTR element indicated under each panel; the fifth panel shows 10 μg *Branchiostoma floridae* (amphioxus) genomic DNA digested with *Hpa*II (H), *Mbo*I (Mb), *Msp*I (M) or *Sau*3A (S) and probed with BfCR1.

RNH. The phylogenetic analysis of the reverse transcriptase domain together with the other structural hallmarks improved the assignment of the *Ciona* sequence to the LOA clade (bootstrap 100%). Finally, the phylogenetic analysis and structural features clearly placed CiR2 within the R2 clade (Figures 1-3). As well as the reverse transcriptase, the preservation in the deuterostome lineage of the distinctive R2 structural hallmarks, such as the REL-endo domain and the 5' CCHH and c-Myb DNA-binding motifs, indicate the ancient structural organization of this clade. Additionally, insertions of the element near the *Ciona* rRNA genes suggest that target specificity through the REL-endo mechanism has been preserved. Overall, not only does the analysis of CiI, CiLOA and CiR2 agree with the origin of these retrotransposons in the Precambrian era [11], but the fact that the urochordate elements resemble the protostome counterparts points to their ancient structural organization.

We derived CiL1 and CiL1.2, whose structural organization and phylogenetic relationship made them cluster within the L1 clade (bootstrap 70%) and supported a previous BLAST analysis of two short *Ciona* sequences [3]. Our data allowed the first structural characterization of the L1 clade in invertebrates. CiL2 clustered within the L2 clade (bootstrap 97%), a novel group of non-LTR retrotransposons closely related to the CR1 and Rex1 clades [14], which includes members previously described in the protostome and deuterostome lineages (Table 2). Interestingly, the CR1 and RTE clades, which are also shared by protostomes and deuterostomes, have not been identified in *Ciona*. Whether these clades were lost in the whole urochordate subphylum needs further investigation.

### Retrotransposon frequency, genomic features and genome evolution

Sequence analysis of the scaffolds harboring the non-LTR elements revealed that ascidian transposable elements are flanked by regions of low gene density. However, no differences in GC content with respect to the average genome value were found when comparing these genomic segments. Moreover, Southern analysis showed that ascidian non-LTR retrotransposons are unmethylated. Overall, these data suggest

**Table 1**

**Copy number of the *Ciona* non-LTR elements**

| | Total number* | Number based on reverse transcriptase | | | |
|---|---|---|---|---|---|
| | | Screening | Slot-blot | Databank | Average |
| CiI | 9 | 7.5 | 5 | 6 | 6 |
| CiL1 | 22 | 3 | 3 | 6 | 4 |
| CiL2 | 24 | 3 | 5 | 4 | 4 |
| CiLOA | 69 | 6 | 5 | 9 | 7 |
| CiR2 | 13 | 3 | - | 3 | 3 |

**In silico* estimates based on all the sequence available.

that mobile elements, gene density and methylation status have not influenced the nucleotide composition in urochordates.

Copy-number estimates of the non-LTR elements in the *Ciona* genome suffer from slight inaccuracies due to the hybridization reaction, which disregards highly divergent elements, and to the fact that computational estimates only refer to the available 90% of the genome. However, the agreement between the *in silico* and experimental estimations indicates that, in this case, the biases have been minimized. The data show a low copy number per haploid genome of the different ascidian elements: from 9 to 69 copies, which decreased to three to seven copies when estimates were based on the reverse transcriptase domain only. These values are far below the vertebrate counterparts, but similar to numbers reported for protostome genomes. This also seems to apply to another lower chordate genome. In amphioxus (subphylum Cephalochordata) a low copy number has been reported for a non-LTR retrotransposon, BfCR1, [16] and for ATE1, a class II transposable element [19].

The factors involved in retrotransposon control are still an open question. The view that methylation evolved to suppress the activity of transposable elements in vertebrates [20] pointed to DNA methylation as a good candidate for transposition control in lower chordates. However, ascidian transposable elements are clearly unmethylated ([21] and this study) and, hence, the genome-defense model cannot be extended to urochordates, and perhaps not to cephalochordates, as the amphioxus BfCR1 element also belongs to the unmethylated genomic fraction (Figure 4c). Among other mechanisms, if required at all for retrotransposon control in lower chordates, co-suppression, which operates on I elements in *Drosophila* [22,23] and in transposable element silencing in plant genomes [24], is a possibility.

## Conclusions

In summary, ascidian and amphioxus genomes do not harbor high copy numbers of retrotransposons. If this reflects the condition of the pre-duplicative genome of the ancestor of the vertebrates, substantial increases in the number of transposons in vertebrates could only have been attained after the large-scale duplications that provided the raw material to buffer the transposable element-induced genome rearrangements, and after the recruitment of methylation to control transposable element mobility. Therefore, beyond the extensive duplications occurring at the origin of the vertebrates, expansion of mobile elements linked to new roles for DNA methylation would have to be considered as significant factors in the modeling of the highly complex genomes.

## Materials and methods

### Non-LTR retrotransposons in the *Ciona* database

The *C. intestinalis* non-LTRs were identified through a TBLASTX [25] search on the *Ciona* genome draft deposited in the JGI database [26]. The following non-LTR retrotransposons were used as queries: CRE1 from *Crithidia fasciculata* (accession number M33009), CZAR from *Trypanosoma cruzi* (M62862), Dong from *Bombyx mori* (L08889), L1 from *Rattus norvegicus* (U83119), RTE1 from *Caenorhabditis elegans* (AF025462), Tad1 from *Neurospora* (L25662), R1 from *D. melanogaster* (X51968), Jockey from *D. melanogaster* (M22874), L1Tc from *T. cruzi* (X83098), R2 from *Porcellio scaber* (AF015818), LOA from *Drosophila silvestris* (X60177), Rex3 from *Tetraodon nigroviridis* (AJ312226), NeSL-1 from *C. elegans* (Z82058), CR1 from *Gallus gallus* (AAC60281) and Maui from *Takifugu rubripes* (AF086712). The retrieved *Ciona* sequences were aligned by eye on the basis of the DotPlot comparisons of the MegAlign program from the DNASTAR package, and a consensus composite was assembled. Sequence differences between scaffolds due to nucleotide substitutions or indels were analyzed and the sequence maximizing the similarity to reported elements was selected. The non-LTR nature of each composite sequence was further verified through a TBLASTX search against the GenBank database. The consensus sequence was named after the defined non-LTR clade to which it belonged. The CiI consensus sequence was derived from scaffolds 120, 148, 599, 1398 and 2116; CiL1 from 951, 1407, 1810, 3249 and 7743;

**Table 2**

**Non-LTR retrotransposons in *Ciona***

| | Deuterostomes | | Protostomes | Other organisms‡ |
|---|---|---|---|---|
| Non-LTR clade | *Ciona* copy number* | Vertebrates (copy number in *Fugu* [15]) | Copy number in *D. melanogaster* [30] and *A. gambiae* [31]† | |
| CR1 | | + (NF§) | + (NF, 152) | |
| CRE | | | + (NF, NF) | |
| I | + (6-9) | | + (67, 19) | |
| Jockey | | | + (392, 28) | |
| L1 | + (4-22) | + (500) | +¶ | + |
| L2 | + (4-24) | + (6,500) | + (NF, NF) | |
| LOA | + (7-69) | | + (18, 19) | |
| NeSL1 | | | + (NF, NF) | |
| R1 | | | + (130, 3¥) | |
| R2 | + (3-13) | | + (0, NF) | |
| R4 | | + (1,000) | + (NF, 2) | |
| Rex1 | | + (2,000) | | + |
| RTE | | + (2,300) | + (NF, 167) | |
| Tad1 | | | | + |
| Total copy number | 24-137 | 12,300 | 607, 390 | |
| Clade complexity** | Five in *Ciona* | Five in *Fugu* | Five in *D. melanogaster* and seven in *A. gambiae* | |

*Detailed in Table 1. †The first number in parentheses refers to *D. melanogaster*, the second to *Aedes aegypti*. ‡See Figure 3. §NF, not found. ¶Probably present in *A. gambiae* [31]. ¥Underestimated in Holt *et al.* [31]. **Number of different clades.

CiL1.2 from 388, 890, 1138, 2278 and 2648; CiL2 from 231, 604, 1005, 1177, 1644 and 3322; CiLOA from 398, 925, 1078, 1854 and 4983; CiR2 from 345, 1777, 2388, 2455 and 3439.

The *in silico* copy number of each repetitive DNA element was estimated from the *Ciona* database. Two types of search were performed. First, all the derived sequences were used to retrieve scaffold-containing elements that matched with a BLAST expect value of $<10^{-3}$ [15]. To discard wrongly assigned elements, a threshold was defined at the score value of the first match of an element that belonged to another clade. Second, for the sake of comparison with experimental data, only the consensus reverse transcriptase region was used for the search and the scaffolds showing a minimum match of 300 nucleotides with the same BLAST expect value were considered.

## PCR amplification, cloning and sequence of non-LTR elements

PCR amplifications with primers designed from the consensus reverse transcriptase sequence of each identified ascidian non-LTR elements were performed with 250 pg of genomic DNA and 1 U Taq DNA polymerase (BioTherm) in 25 μl of reaction volume containing 0.2 μM for each primer, 32 μM each dNTP and 2 mM $MgCl_2$. The sequences of the primers were: CiL1-F (forward): 5'-AACTAGTGATACCGCGCC-3', CiL1-R (reverse): 5'-ACACCTCGTTTGATCGG-3', CiL2-F: 5'-GTTGAGGTAAATGGCGC-3', CiL2-R: 5'-CGTTCGTCATTATCTGGG-3', CiR2-F: 5'-TTCCGCAAGGTCGATG-3', CiR2-R: 5'-CAGATAGGGCCCAATCC-3', CiI-F: 5'-CGATCTACCACCGACCAC-3', CiI-R: 5'-GCTTGTCACAGGCAGTTG-3', CiLOA-F: 5'-AACTGCGGAGATCCATGG-3' and CiLOA-R: 5'-GTCGCAGTCTTGATGCGG-3'. PCR conditions were as follows: the initial denaturation step at 94°C for 2 min was followed by 40 cycles at 94°C for 45 sec, 53°C for 30 sec and 72°C for 30 sec and a final extension step at 72°C for 5 min. In each PCR assay, a fragment of approximately 300 bp was amplified and then cloned in a pUC18 plasmid and sequenced using the Big Dye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA) in a 3730 DNA Analyzer (Applied Biosystems).

## Genomic library screenings, slot and Southern blot analyses

A *C. intestinalis* λZapII genomic library (kindly provided by M. Levine) was screened with each of the fragments of the identified *Ciona* non-LTR elements. The probes were labeled with [$\alpha$-$^{32}$P]dCTP by random-hexamer priming and hybridized to phage DNA transferred on Hybond-N nylon filters (Amersham Pharmacia Biotech, Uppsala, Sweden) in duplicate. Approximately 70,000 phages were screened. Hybridizations were performed in phosphate-SDS solution [27] at 65°C overnight. Two 15-min washes were performed at 65°C in 2 × SSC, 0.1% SDS, 2 × 15 min at 65°C in 1 × SSC, 0.1% SDS

and 1 × 15 min at 65°C in 0.2 × SSC, 0.1% SDS. Hybridization signals were detected by autoradiography. Only the signals present in the original and duplicated filters were considered.

For quantitative slot-blot analysis, 500 ng, 250 ng, 50 ng and 25 ng of *Eco*RI-digested *C. intestinalis* genomic DNA and serial dilutions of each plasmid-containing probe, *Eco*RI-restricted and mixed with 1 μg mouse genomic DNA (as non-specific DNA) were denatured with 0.4 M NaOH and 25 mM EDTA in a final volume of 200 μl and blotted on Hybond-N nylon filters (Amersham Pharmacia Biotech) with a slot-blot device (Minifold II, Schleicher & Schuell, Dassel, Germany). Three genomic DNA replicates of isolated animals were performed. Before sample loading, the membrane was soaked in water and then neutralized with 2 M sodium acetate, pH 5.4 and fixed with UV light. Membranes were hybridized with the same probes used for library screening at the same hybridization and washing conditions. The slot-blot signal was quantified with the GS525 Molecular Imager System (Bio-Rad, Hercules, CA).

For Southern analyses 10 μg of *C. intestinalis* genomic DNA digested with *Hpa*II or *Msp*I was resolved on 0.8% agarose gels and transferred to nylon membranes. Southern blots were hybridized with the same non-LTR probes used for library screening at identical hybridization and washing conditions.

### Sequence and phylogenetic analyses

Gene density and GC content of the retrotransposon insertion sites was assessed. Only the scaffolds that expanded at least 10 kb upstream or downstream from an element were considered. Gene density in the 10-kb flanking regions was calculated by scoring the predicted genes according the *Ciona* gene model v1.0. When the gene sequence was only partially contained in the region analyzed, it was scored as 0.5.

For phylogenetic analysis the *C. intestinalis* sequences were added to a previous alignment by Malik [11], updated by adding the NeSL-1 (*C. elegans*, Z82058), LINE2 (*Patella*, X77618; *Danio rerio*, AL591210; *Oryzias*, AB054295; *Fugu*, AF086712) and Rex1 (*Xiphophorus*, AF155728; *Batrachocottus*, AAA83744) clades. The new alignment was generated using Clustal X [28], maintaining the same pairwise gap penalties and multiple alignment parameters, and adjusted by eye (see Additional data files 2 and 3). Phylogenetic analyses were performed using the neighbor-joining method, rooted with the reverse transcriptase sequence of *Neurospora* organellar group II intron (accession number S07649) and drawn with the TreeViewPPC program [29]. Confidence in each node was assessed by 1,000 bootstrap replicates.

## Additional data files

The consensus DNA sequences of each derived *Ciona* non-LTR retrotransposon (Additional data file 1) and the reverse transcriptase alignment used to reconstruct the phylogenetic relationship with the non-LTR clades (Additional data file 2, Additional data file 3) are available with the online version of this article.

## Acknowledgements

We thank G. Marfany for helpful discussion and R. Rycroft for revising the English. We also thank the Serveis Científico-Tècnics (UB) for DNA sequencing. This study was supported by the Ministerio de Educación y Cultura (grant BMC2000-0536 and BMC 2003-05211). J.P. was the recipient of an fellowship from the Universitat de Barcelona.

## References

1. Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM, *et al.*: **The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins.** *Science* 2002, **298:**2157-2167.
2. Schmidtke J, Epplen JT, Engel W: **Genome analysis of Amphioxus and speculation as to the origin of contrasting vertebrate genome organization patterns.** *Comp Biochem Physiol* 1979, **63:**455-458.
3. Simmen MW, Bird A: **Sequence analysis of transposable elements in the sea squirt, Ciona intestinalis.** *Mol Biol Evol* 2000, **17:**1685-1694.
4. Charlesworth B, Sniegowski P, Stephan W: **The evolutionary dynamics of repetitive DNA in eukaryotes.** *Nature* 1994, **371:**215-220.
5. Finnegan DJ: **Eukaryotic transposable elements and genome evolution.** *Trends Genet* 1989, **5:**103-107.
6. Finnegan DJ: **Transposable elements.** *Curr Opin Genet Dev* 1992, **2:**861-867.
7. Xiong Y, Eickbush TH: **Origin and evolution of retroelements based upon their reverse transcriptase sequences.** *EMBO J* 1990, **9:**3353-3362.
8. Luan DD, Korman MH, Jakubczak JL, Eickbush TH: **Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition.** *Cell* 1993, **72:**595-605.
9. Burch JB, Davis DL, Haas NB: **Chicken repeat 1 elements contain a pol-like open reading frame and belong to the non-long terminal repeat class of retrotransposons.** *Proc Natl Acad Sci USA* 1993, **90:**8199-8203.
10. Martín F, Olivares M, López MC: **Do non-long terminal repeat retrotransposons have nuclease activity?** *Trends Biochem Sci* 1996, **21:**283-285.
11. Malik HS, Burke WD, Eickbush TH: **The age and evolution of non-LTR retrotransposable elements.** *Mol Biol Evol* 1999, **16:**793-805.
12. Malik HS, Eickbush TH: **NeSL-1, an ancient lineage of site-specific non-LTR retrotransposons from *Caenorhabditis elegans*.** *Genetics* 2000, **154:**193-203.
13. Volff JN, Korting C, Schartl M: **Multiple lineages of the non-LTR retrotransposon Rex1 with varying success in invading fish genomes.** *Mol Biol Evol* 2000, **17:**1673-1684.
14. Lovsin N, Gubensek F, Kordi D: **Evolutionary dynamics in a novel L2 clade of non-LTR retrotransposons in Deuterostomia.** *Mol Biol Evol* 2001, **18:**2213-2224.
15. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, *et al.*: **Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*.** *Science* 2002, **297:**1301-1310.
16. Albalat R, Permanyer J, Cañestro C, Martínez-Mir A, Gonzàlez-Angulo O, Gonzàlez-Duarte R: **The first non-LTR retrotransposon characterised in the cephalochordate amphioxus, BfCR1, shows similarities to CR1-like elements.** *Cell Mol Life Sci* 2003, **60:**803-809.
17. Tu Z, Isoe J, Guzova JA: **Structural, genomic, and phylogenetic analysis of Lian, a novel family of non-LTR retrotransposons in the yellow fever mosquito, *Aedes aegypti*.** *Mol Biol Evol* 1998, **15:**837-853.
18. Burke WD, Malik HS, Jones JP, Eickbush TH: **The domain structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods.** *Mol Biol Evol* 1999, **16:**502-511.

19. Cañestro C, Albalat R, Gonzàlez-Duarte R: **Isolation and characterization of the first non-autonomous transposable element in amphioxus, ATE-1.** *Gene*, in press.
20. Yoder JA, Walsh CP, Bestor TH: **Cytosine methylation and the ecology of intragenomic parasites.** *Trends Genet* 1997, **13:**335-340.
21. Simmen MW, Leitgeb S, Charlton J, Jones SJ, Harris BR, Clark VH, Bird A: **Nonmethylated transposable elements and methylated genes in a chordate genome.** *Science* 1999, **283:**1164-1167.
22. Chaboissier MC, Bucheton A, Finnegan DJ: **Copy number control of a transposable element, the I factor, a LINE-like element in *Drosophila*.** *Proc Natl Acad Sci USA* 1998, **95:**11781-11785.
23. Jensen S, Gassama MP, Heidmann T: **Taming of transposable elements by homology-dependent gene silencing.** *Nat Genet* 1999, **21:**209-212.
24. Flavell RB: **Inactivation of gene expression in plants as a consequence of specific sequence duplication.** *Proc Natl Acad Sci USA* 1994, **91:**3490-3496.
25. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
26. **DOE Joint Genome Institute** [http://www.jgi.doe.gov]
27. Church GM, Gilbert W: **Genomic sequencing.** *Proc Natl Acad Sci USA* 1984, **81:**1991-1995.
28. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Res* 1997, **25:**4876-4882.
29. Page RDM: **TREEVIEW: An application to display phylogenetic trees on personal computers.** *Comput Appl Biosci* 1996, **12:**357-358.
30. Berezikov E, Bucheton A, Busseau I: **A search for reverse transcriptase-coding sequences reveals new non-LTR retrotransposons in the genome of *Drosophila melanogaster*.** *Genome Biol* 2000, **1:**research0012.1-0012.15.
31. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JM, Wides R, *et al.*: **The genome sequence of the malaria mosquito *Anopheles gambiae*.** *Science* 2002, **298:**129-149.
32. Fawcett DH, Lister CK, Kellett E, Finnegan DJ: **Transposable elements controlling I-R hybrid dysgenesis in *D. melanogaster* are similar to mammalian LINEs.** *Cell* 1986, **47:**1007-1015.

**Capítol 2: Retrotransposons non-LTR a *Branchiostoma floridae***

Article 2: *The first non-LTR retrotransposon characterised in the cephalochordate amphioxus, BfCR1, shows similarities to CR1-like elements.* **Cell Mol Life Sci. 2003 Apr;60(4):803-9**

Article 3: *Getting closer to a pre-vertebrate genome: the non-LTR retrotransposons of* Branchiostoma floridae. **Int J Biol Sci. 2006;2(2):48-53**

Els retrotransposons non-LTR de diferents espècies de cordats presenten situacions diverses. Donat que l'amfiox s'havia considerat històricament com el cordat no vertebrat més proper als vertebrats era especialment interessant estudiar aquests elements mòbils en el seu genoma. En el primer treball (article 2) es caracteritza el primer retrotransposó non-LTR de *Branchiostoma floridae,* anomenat BfCR1. Aquest element presenta unes característiques més similars a les descrites per als retrotransposons non-LTR de l'ascidi *Ciona intestinalis* que les dels vertebrats. L'anàlisi de la seqüència de BfCR1 mostra clarament que és un element de tipus CR1 amb els dominis característics d'aquest clade. Tal com succeeix a l'ascidi, el número de còpies de BfCR1 en el genoma és baix, recolzant la idea que els cordats no vertebrats, organismes amb genomes relativament "senzills i de mida reduïda" contenen pocs retroposons non-LTR en el seu genoma.

En el segon treball (article 3) i gràcies a la base de dades de la seqüència del genoma de l'amfiox, va ser posible realitzar una anàlisi molt més general dels retrotransposons non-LTR en aquest organisme. El crivellatge *in silico* que hem dut a terme de les seqüències genèomiques ens ha permès establir la presència de retrotransposons non-LTR corresponents a almenys 6 clades diferents. Tot i que no hem pogut caracteritzar l'entorn genòmic d'aquests elements l'estima del nombre de còpies és clarament baix. Aquest treball ha permès generalitzar les dades experimentals de l'element BfCR1 a altres tipus de retrotransposons dins l'amfiox, fer una anàlisi comparativa amb els elements descrits en d'altres espècies animals i, finalment, reafirmar i extendre les nostres conclusions anteriors sobre les característiques dels TEs als cordats més primitius.

Aquesta pàgina correspon a l'article "The first non-LTR retrotransposon characterised in the cephalochordate amphioxus, BfCR1, shows similarities to CR1-like elements" publicat a la revista *Cellular and Molecular Life Sciences*. Ja que aquesta revista posseeix els drets de distribució i publicació, s'omet la pàgina original. Els resultats d'aquest article són ampliament explicats i discutits en les corresponents seccions d'aquesta memòria.

Aquesta pàgina correspon a l'article "The first non-LTR retrotransposon characterised in the cephalochordate amphioxus, BfCR1, shows similarities to CR1-like elements" publicat a la revista *Cellular and Molecular Life Sciences*. Ja que aquesta revista posseeix els drets de distribució i publicació, s'omet la pàgina original. Els resultats d'aquest article són ampliament explicats i discutits en les corresponents seccions d'aquesta memòria.

Aquesta pàgina correspon a l'article "The first non-LTR retrotransposon characterised in the cephalochordate amphioxus, BfCR1, shows similarities to CR1-like elements" publicat a la revista *Cellular and Molecular Life Sciences*. Ja que aquesta revista posseeix els drets de distribució i publicació, s'omet la pàgina original. Els resultats d'aquest article són ampliament explicats i discutits en les corresponents seccions d'aquesta memòria.

Aquesta pàgina correspon a l'article "The first non-LTR retrotransposon characterised in the cephalochordate amphioxus, BfCR1, shows similarities to CR1-like elements" publicat a la revista *Cellular and Molecular Life Sciences*. Ja que aquesta revista posseeix els drets de distribució i publicació, s'omet la pàgina original. Els resultats d'aquest article són ampliament explicats i discutits en les corresponents seccions d'aquesta memòria.

Aquesta pàgina correspon a l'article "The first non-LTR retrotransposon characterised in the cephalochordate amphioxus, BfCR1, shows similarities to CR1-like elements" publicat a la revista *Cellular and Molecular Life Sciences*. Ja que aquesta revista posseeix els drets de distribució i publicació, s'omet la pàgina original. Els resultats d'aquest article són ampliament explicats i discutits en les corresponents seccions d'aquesta memòria.

Aquesta pàgina correspon a l'article "The first non-LTR retrotransposon characterised in the cephalochordate amphioxus, BfCR1, shows similarities to CR1-like elements" publicat a la revista *Cellular and Molecular Life Sciences*. Ja que aquesta revista posseeix els drets de distribució i publicació, s'omet la pàgina original. Els resultats d'aquest article són ampliament explicats i discutits en les corresponents seccions d'aquesta memòria.

Aquesta pàgina correspon a l'article "The first non-LTR retrotransposon characterised in the cephalochordate amphioxus, BfCR1, shows similarities to CR1-like elements" publicat a la revista *Cellular and Molecular Life Sciences*. Ja que aquesta revista posseeix els drets de distribució i publicació, s'omet la pàgina original. Els resultats d'aquest article són ampliament explicats i discutits en les corresponents seccions d'aquesta memòria.

**International Journal of Biological Sciences**
ISSN 1449-2288 www.biolsci.org 2006 2(2):48-53


**Research paper**

# Getting closer to a pre-vertebrate genome: the non-LTR retrotransposons of *Branchiostoma floridae*

**Jon Permanyer, Ricard Albalat and Roser Gonzàlez-Duarte**

Departament de Genètica. Facultat de Biologia. Universitat de Barcelona. 08028 Barcelona, Spain.

Corresponding address: Roser Gonzàlez-Duarte, Departament de Genètica, Facultat de Biologia, Universitat de Barcelona. Av. Diagonal, 645. 08028 Barcelona, Spain. Tel.: +34.934021034; Fax: +34.934034420; E-mail: rgonzalez@ub.edu

Received: 2006.02.06; Accepted: 2006.03.10; Published: 2006.04.10

Non-LTR retrotransposons are common in vertebrate genomes and although present in invertebrates they appear at a much lower frequency. The cephalochordate amphioxus is the closest living relative to vertebrates and has been considered a good model for comparative analyses of genome expansions during vertebrate evolution. With the aim to assess the involvement of transposable elements in these events, we have analysed the non-LTR retrotransposons of *Branchiostoma floridae*. In silico searches have allowed to reconstruct non-LTR elements of six different clades (CR1, I, L1, L2, NeSL and RTE) and assess their structural features. According to the estimated copy number of these elements they account for less than 1% of the haploid genome, which reminds of the low abundance also encountered in the urochordate *Ciona intestinalis*. Amphioxus (*B. floridae*) and *Ciona* share a pre-vertebrate-like organization for the non-LTR retrotransposons (<150 copies, < 1% of the genome) versus the complexity associated to higher vertebrates (*Homo sapiens* >1.3·10$^6$ copies, > 20% of the genome).

Key words: transposable elements, non-LTR retrotransposons, cephalochordates, genome evolution.

## 1. Introduction

Transposable elements (TEs) are almost invariably found in all species that have been studied. TEs are classified according to their degree of self-sufficiency and to their mechanism of transposition [1]. Regarding the first, TEs are divided in autonomous and nonautonomous elements. Based on the mode of transposition, two classes of TEs are defined: class I elements or retroelements (which utilize reverse transcription to amplify) and class II or DNA transposons (which transpose by the cut-and-paste or the rolling circle mode). This work has focussed on the autonomous class I elements non-LTR retrotransposons (also called LINE-like elements, polyA retrotransposons or retroposons) of the cephalochordate *Branchiostoma floridae*.

Non-LTR retrotransposons are one of the most abundant classes of transposable elements that make up a substantial fraction of the vertebrate genome. They comprise a variety of dispersed sequences that cluster in at least 14 clades and are divided in two groups, old-LINEs or site-specific endonuclease retrotransposons encoded in a single open reading frame (ORF), and young-LINEs or non-site-specific endonuclease retrotransposons that encode two ORFs (ORF1 and ORF2) [1, 2]. Both groups codify a preserved reverse transcriptase (RT), the only common domain, strictly required to achieve transposition and frequently used to analyse phylogenetic relationships. Additional structural motives are, a restriction enzyme-like endonuclease (REL-endo) or an apurinic/apyrimidinic endonuclease (APE), of those, at least one is strictly required and, optionally, several nucleic acid binding domains (NABD) and an RNAse H signature. Irrespective of the type of non-LTR retrotransposons, overall copy number is high enough not to leave them aside when dealing with genome evolution. Regarding TEs in general, their contribution to genome rearrangements has been deeply reported (reviewed in [1]).

Amphioxus (*B. floridae*) is a key organism to understand the invertebrate to vertebrate transition because it possesses a prototypical chordate body plan and is considered the closest living relative to vertebrates. The genome of this animal is small and relatively unduplicated, as shown by the single cluster of 14 Hox genes vs the four, or even more, clusters described in vertebrates [reviewed in 3]. Moreover, the recent availability of the genome draft of the amphioxus *B. floridae* has facilitated the analysis and comparison of non-LTR retrotransposons with those of the urochordate *Ciona intestinalis* and other vertebrate species.

## 2. Materials and methods

### In silico search of non-LTR retrotransposons

The *Branchiostoma floridae* non-LTR elements were identified through a local TBLASTN [4] search of the first 4,772,554 *B. floridae* whole genome shotgun sequences (8xcoverage) generated at the JGI (www.jgi.doe.gov) and deposited in the Ensemble traces database (ftp.ensembl.org/pub/traces/

branchiostoma_floridae). The following sequences were used as queries: CRE1 and CRE2 from *Crithidia fasciculata* (accession numbers M33009 and U19151), CZAR from *Trypanosoma cruzi* (M62862), Slacs from *Trypanosoma brucei* (X17078), Dong from *Bombyx mori* (L08889), R4Pe from *Parascaris equorum* (U31672), L1 from *Rattus norvegicus* (U83119), Zepp from *Chlorella vulgaris* (AB008896), Tx1L from *Xenopus laevis* (M26915), RTE1 from *Caenorhabditis elegans* (AF025462), Bov-B from *Vipera ammodytes* (AF332697), Rex3 from *Tetraodon nigroviridis* (AJ312226), Tad1 from *Neurospora* (L25662), Mgr583 from *Magnaporthe grisea* (AF018033), R1 from *Drosophila melanogaster* (X51968), RT1 from *Anopheles gambiae* (M93690), Jockey from *D. melanogaster* (M22874), Helena from *D. mercatorum* (AF015277), JuanC from *Culex pipiens* (M91082), L1Tc from *T. cruzi* (X83098), Idt from *D. teisseri* (M28878), R2 from *Porcellio scaber* (AF015818), R2 from *Forficula auricularia* (AF015819), LOA from *D. silvestris* (X60177), Trim from *D. miranda* (X59239), Bilbo-1 from *D. subobscura* (U73800), NeSL-1 from *C. elegans* (Z82058 and NM_075007), Rex1 from *Batrachocottus baicalensis* (AAA83744), CR1 from *Gallus gallus* (AAC60281), *Bf*CR1 from *B. floridae* (AF369890), T1 from *A. gambiae* (M93689), Sam6 from *C. elegans* (U46668) and Maui from *Takifugu rubripes* (AF086712). Overlapping clones, identified through local BLASTN searches, were used to walk in silico upstream and downstream of each sequence. For every element identified, consensus nucleotide sequence were assembled from all the overlapping clones with an expected value of $<10^{-200}$ with the Seqman II software [5], wich usually generates only one composite with some ambiguities and TGI Clustering Tools software (www.tigr.org) with an strict algorithm which generates more than one composite with no ambiguities. Only the assemblies composed from more than 10 sequences were considered. The non-LTR nature of each composite sequence was further verified by reciprocal best BLAST search against the GenBank database. The consensus sequence was named after the defined non-LTR clade to which it belonged.

### Copy number

The copy number for each non-LTR retrotransposon per haploid genome was determined as described [6], by multiplying the number of matching shotgun clones with an expected value of $<10^{-200}$ by $5.8 \cdot 10^{8}$bp the size of the *B. floridae* haploid genome and divided by the length of the composite and the number of shotgun sequences in the local database (4,772,554).

### Phylogenetic analysis

The RT deduced sequences of *B. floridae* were added to a previous alignment [7] and a new one was generated with Clustal X [8], maintaining the same pairwise gap penalties and multiple alignment parameters (Fig 1). Phylogenetic analyses were performed using the neighbor-joining method, rooted with the *Neurospora* organellar group II intron (accession number S07649) and drawn with the TreeViewPPC program [9]. Confidence in each node was assessed by 1,000 bootstrap replicates.

## 3. Results

We have screened the non-LTR retrotransposons in the shotgun genome project of *B. floridae* in order to characterise the type and number of elements and draw a comparison with other known genomes.

Searches identified members of six out of fourteen previously reported clades. According to the phylogeny established and the Genbank comparisons, they will be termed *Bf*CR1, *Bf*I, *Bf*L1, *Bf*L2, *Bf*NeSL and *Bf*RTE. Comparisons of the composites of each element allowed to define two conserved domains: RT and APE. The RT domain, described in all amphioxus clades, contained all the distinctive structural hallmarks defined as block 0, 1, 2, 2a, 3, 4, 5, 6, 7, 8 and 9 [10]. Moreover, the apurinic/apyrimidinic endonuclease (APE) region was identified with reasonable confidence in *Bf*CR1, *Bf*L1, *Bf*L2 and *Bf*RTE (Fig. 1) on the basis of the reported domains I to IX [11] and only the last domain in *Bf*I, which supports the *bona fide* structure of the defined composite and argues against a non-TE-based assembly. Notwithstanding our exhaustive search, the N-terminal APE domain and the RNaseH (RNH) sequence were not detected in *Bf*I elements; neither REL-endo signatures could be clearly characterised in *Bf*NeSL. For none of the elements identified, either NABD or ORF1 sequences could be detected. Copy number of each element per haploid genome was determined from the whole-length available sequence. In silico estimates showed low copy numbers: 25 for *Bf*CR1, 3 for *Bf*I, 32 for *Bf*L1, 35 for *Bf*L2, 6 for *Bf*NeSL and 42 for *Bf*RTE (Table 1). A rough estimation of the genome fraction harbouring non-LTR retrotransposons could be obtained considering that all the estimated copies (143) correspond to 5 kb full-length elements, and the value obtained would represent less than 1% of the haploid genome.

Intra-sequence variability for each of the 6 clades was assessed from amino acid sequence comparison of the RT domains and expressed by the degree of similarity in percentage (Fig 2C).

The matrices gave a range of 32.5-98.1% for *Bf*CR1, 53.4-93.6% for *Bf*I, 23.9-97.6% for *Bf*L1, 22.4-99.5% for *Bf*L2, 31.7-85.4% for *Bf*NeSL and 15.4-91.7% for *Bf*RTE.

**Figure 1**. Alignment of the deduced protein sequence of the consensus contig of each clade. APE domains I, II, III, V, VI, VIII and IX, and the RT structural blocks 0-9 are indicated. Amino acid identities and similarities are shown in black and gray shading, respectively

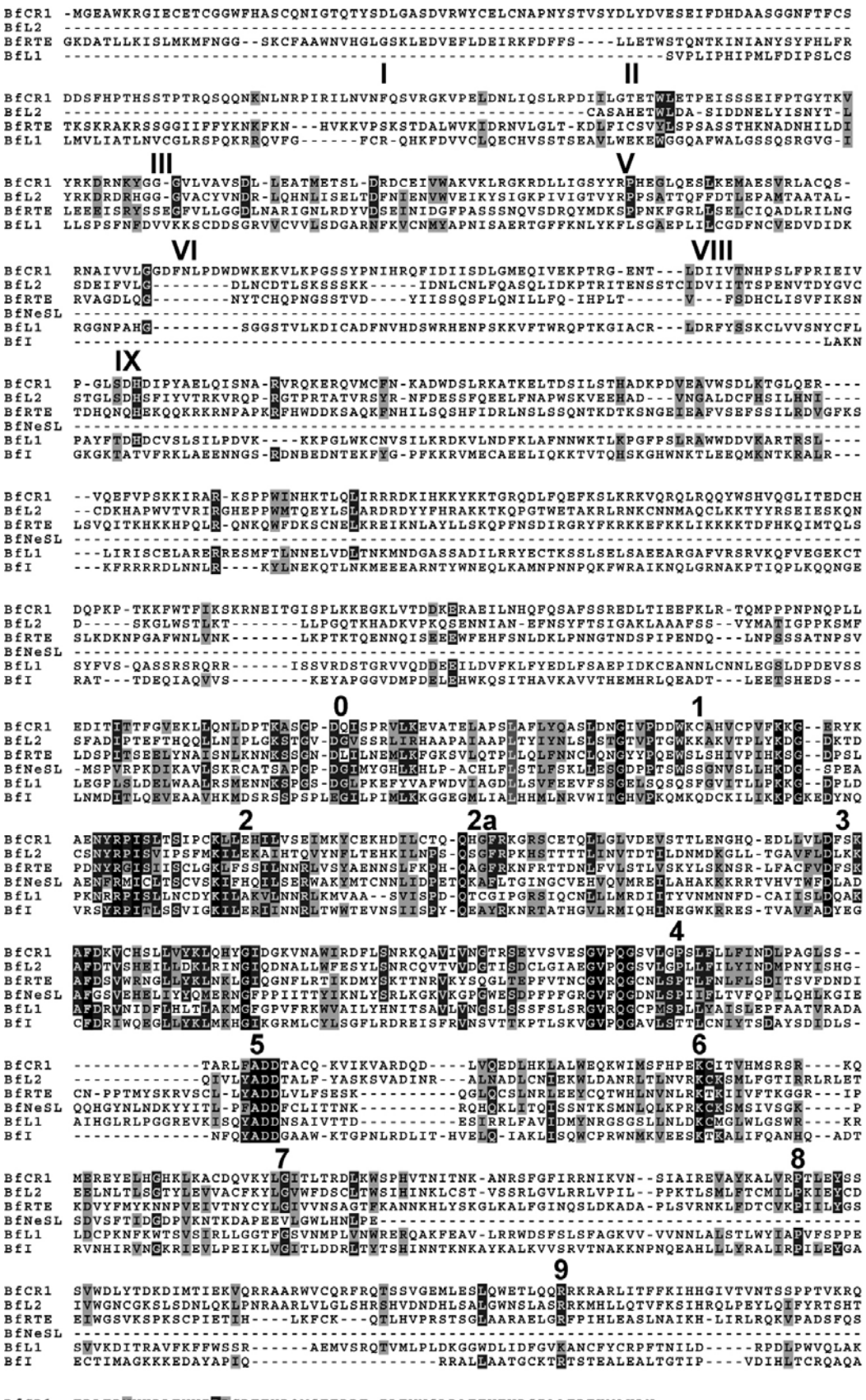
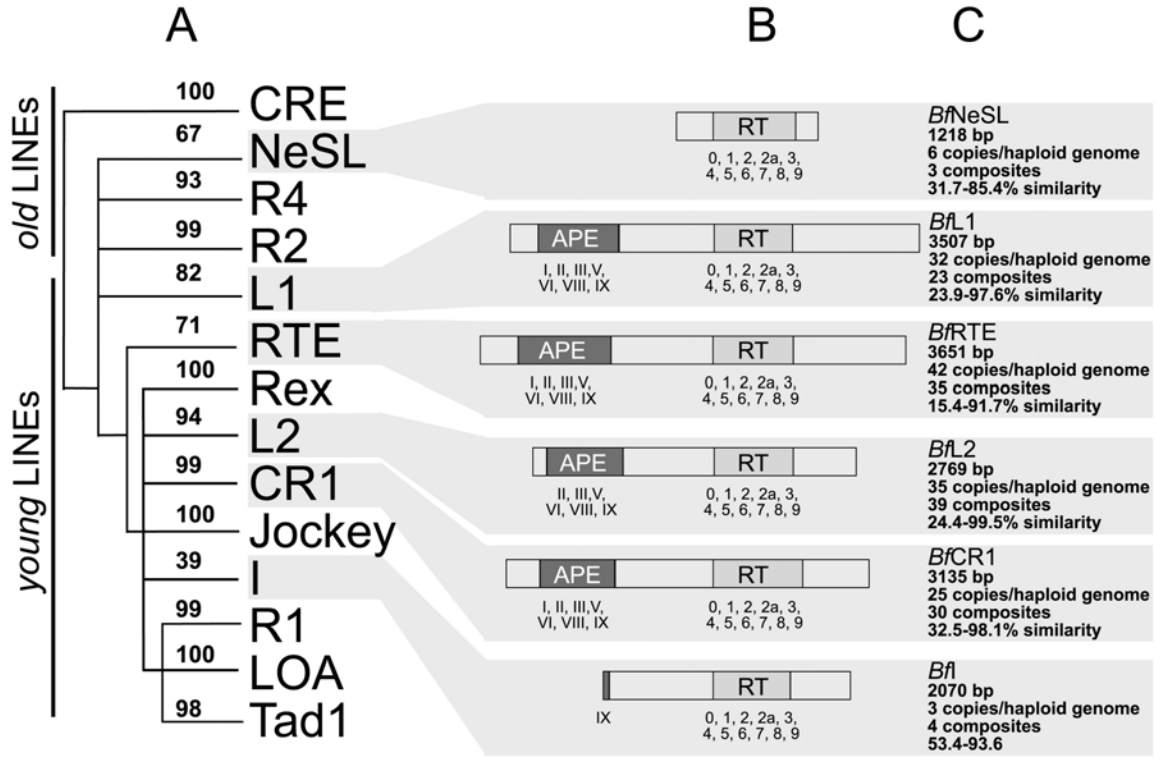

BfCR1 -MGEAWKRGIECETCGGWFHASCQNIGTQTYSDLGASDVRWYCELCNAPNYSTVSYDLYDVESEIFDHDAASGGNFTFCS
BfL2 ----------
BfRTE GKDATLLKISLMKMFNGG--SKCFAAWNVHGLGSKLEDVEFLDEIRKFDFFS----LLETWSTQNTKINIANYSYFHLFR
BfL1 ----------SVPLIPHIPMLFDIPSLCS
I II
BfCR1 DDSFHPTHSSTPTRQSQQNKNLNRPIRILNVNFQSVRGKVPELDNLIQSLRPDIILGTETWLETPEISSSEIFPTGYTKV
BfL2 ----------CASAHETWLDA-SIDDNELYISNYT-L
BfRTE TKSKRAKRSSGGIIFFYKNKFKN---HVKKVPSKSTDALWVKIDRNVLGLT-KDLFICSVYLSPSASSTHKNADNHILDI
BfL1 LMVLIATLNVCGLRSPQKRRQVFG------FCR-QHKFDVVCLQECHVSSTSEAVLWEKEWGGQAFWALGSSQSRGVG-I
III V
BfCR1 YRKDRNKYGG-GVLVAVSDL-LEATMETSL-DRDCEIVWAKVKLRGKRDLLIGSYYRPHEGLQESLKEMAESVRLACQS-
BfL2 YRKDRDRHGG-GVACYVNDR-LQHNLISELTDFNIENVWVEIKYSIGKPIVIGTVYRPPSATTQFFDTLEPAMTAATAL-
BfRTE LEEEISRYSSEGFVLLGGDLNARIGNLRDYVDSEINIDGFPASSSNQVSDRQYMDKSPPNKFGRLFSELCIQADLRILNG
BfL1 LLSPSFNFDVVKKSCDDSGRVVCVVLSDGARNFKVCNMYAPNISAERTGFFKNLYKFLSGAEPLILCGDFNCVEDVDIDK
VI VIII
BfCR1 RNAIVVLGGDFNLPDWDWKEKVLKPGSSYPNIHRQFIDIISDLGMEQIVEKPTRG-ENT---LDIIVTNHPSLFPRIEIV
BfL2 SDEIFVLG---------DLNCDTLSKSSSKK-----IDNLCNLFQASQLIDKPTRITENSSTCIDVIITTSPENVTDYGVC
BfRTE RVAGDLQG---------NYTCHQPNGSSTVD----YIISSQSFLQNILLFQ-IHPLT------V---FSDHCLISVFIKSN
BfNeSL ----------
BfL1 RGGNPAHG---------SGGSTVLKDICADFNVHDSWRHENPSKKVFTWRQPTKGIACR---LDRFYSSKCLVVSNYCFL
BfI ----------LAKN
IX
BfCR1 P-GLSDHDIPYAELQISNA-RVRQKERQVMCFN-KADWDSLRKATKELTDSILSTHADKPDVEAVWSDLKTGLQER----
BfL2 STGLSDHSFIYVTRKVRQP-RGTPRTATVRSYR-NFDESSFQEELFNAPWSKVEEHAD---VNGALDCFHSILHNI----
BfRTE TDHQNQHEKQQKRKRNPAPKRFHWDDKSAQKFNHILSQSHFIDRLNSLSSQNTKDTKSNGEIEAFVSEFSSILRDVGFKS
BfNeSL ----------
BfL1 PAYFTDHDCVSLSILPDVK----KKPGLWKCNVSILKRDKVLNDFKLAFNNWKTLKPGFPSLRAWWDDVKARTRSL----
BfI GKGKTATVFRKLAEENNGS-RDNBEDNTEKFYG-PFKKRVMECAEELIQKKTVTQHSKGHWNKTLEEQMKNTKRALR---
BfCR1 --VQEFVPSKKIRAR-KSPPWINHKTLQEIRRRDKIHKKYKKTGRQDLFQEFKSLKRKVQRQLRQQYWSHVQGLITEDCH
BfL2 --CDKHAPWVTVRIRGHEPPWMTQEYLSLARDRDYYFHRAKKTKQPGTWETAKRLRNKCNNMAQCLKKTYYRSEIESKQN
BfRTE LSVQITKHKKHPQLR-QNKQWFDKSCNELKREIKNLAYLLSKQPFNSDIRGRYFKRKKEFKKLIKKKKTDFHKQIMTQLS
BfNeSL ----------
BfL1 ---LIRISCELARERRESMFTLNNELVDETNKMNDGASSADILRRYECTKSSLSELSAEEARGAFVRSRVKQFVEGEKCT
BfI ---KFRRRRDLNNLR----KYLNEKQTLNKMEEEARNTYWNEQLKAMNPNNPQKFWRAIKNQLGRNAKPTIQPLKQQNGE
BfCR1 DQPKP-TKKFWTFIKSKRNEITGISPLKKEGKLVTDDKERAEILNHQFQSAFSSREDLTIEEFKLR-TQMPPPNPNQPLL
BfL2 D-----SKGLWSTLKT-------LLPGQTKHADKVPKQSENNIAN-EFNSYFTSIGAKLAAAFSS--VYMATIGPPKSMF
BfRTE SLKDKNPGAFWNLVNK-------LKPTKTQENNQISEEEWFEHFSNLDKLPNNGTNDSPIPENDQ---LNPSSSATNPSV
BfNeSL ----------
BfL1 SYFVS-QASSRSRQRR------ISSVRDSTGRVVQDDEEILDVFKLFYEDLFSAEPIDKCEANNLCNNLEGSLDPDEVSS
BfI RAT---TDEQIAQVVS--------KEYAPGGVDMPDELEHWKQSITHAVKAVVTHEMHRLQEADT---LEETSHEDS---
0 1
BfCR1 EDITITTFGVEKLLQNLDPTKASGP-DQHSPRVLKEVATELAPSLAFLYQASLDNGIVPDDWKCAHVCPVFKKG--ERYK
BfL2 SFADIPTEFTHQQLLNIPLGKSTGV-DGVSSRLIRHAAPAIAAPLTYIYNLSLSTGTVPTGWKKAKVTPLYKDG--DKTD
BfRTE LDSPITSEELYNAISNLKNNKSSGN-DLHLNEMLKFGKSVLQTPLLQLFNNCLQNGYYPQEWSLSHIVPIHKSG--DPSL
BfNeSL -MSPVRPKDIKAVLSKRCATSAPGP-DGIMYGHLKHLP-ACHLFLSTLFSKLLESGDPPTSWSSGNVSLLHKDG--SPEA
BfL1 LEGPLSLDELWAALRSMENNKSPGS-DGLPKEFYVAFWDVIAGDLLSVFEEVFSSGELSQSQSFGVITLLPKKG--DPLD
BfI LNMDITLQEVEAAVHKMDSRSSPSPLEGILPIMLKKGGEGMLIALHHMLNRVWITGHVPKQMKQDCKILIKAPGKEDYNQ
2 2a 3
BfCR1 AENYRPISLTSIPCKLLEHILVSEIMKYCEKHDILCTQ-QHGFRKGRSCETQLLGLVDEVSTTLENGHQ-EDLLVLDFSK
BfL2 CSNYRPISVIPSFMKILEKAIHTQVYNFLTEHKILNPS-QSGFRPKHSTTTTLINVTDTLLDNMDKGLL-TGAVFLDLKK
BfRTE PDNYRGISIISCLGKLFSSILNNRLVSYAENNSLFKPH-QAGFRKNFRTTDNLFVLSTLVSKYLSKNSR-LFACFVDFSK
BfNeSL AENFRMICLTSCVSKIFHQILSERWAKYMTCNNLIDPETQKAGLTGINGCVEHVQVMRELLAHAKKKRRTVHVTWFDLAD
BfL1 PKNRRPISLLNCDYKILAKVLNNRLKMVAA--SVISPD-QTCGIPGRSIQCNLLLMRDILTYVNMNNFD-CAIISLDQAK
BfI VRSYRPITLSSVIGKILERIINNRLTWWTEVNSIISPY-QEAYRKNRTATHGVLRMIQHINEGWKRRES-TVAVFADYEG
4
BfCR1 AFDKVCHSLLVYKLQHYGIDGKVNAWIRDFLSNRKQAVIVNGTRSEYVSVESGVPQGSVLGPSLFLLFINDLPAGLSS--
BfL2 AFDTVSHEILLDKLRINGIQDNALLWFESYLSNRCQVTVVDGTISDCLGIAEGVPQGSVLGPLLFILYINDMPNYISHG-
BfRTE AFDSVWRNGLLYKLNKLGIQGNFLRTIKDMYSKTTNRVKYSQGLTEPFVTNCGVRQGCNLSPTLFNLFLSDITSVFDNDI
BfNeSL AFGSVEHELIYYQMERNGFPPIITTYIKNLYSRLKGKVKGPGWESDPFFFGRGVFQGDNLSPIIFLTVFQPILQHLKGIE
BfL1 AFDRVNIDFWHLTLAKMGFGPVFRKWVAILYHNITSAVLVNGSLSSSFSLSRGVRQGCPMSPLLYAISLEPFAATVRADA
BfI CFDRIWQEGLLYKLMKHGIKGRMLCYLSGFLRDREISFRVNSVTTKPTLSKVGVPQGAVLSTTLCNIYTSDAYSDIDLS-
5 6
BfCR1 -------------TARLFADDTACQ-KVIKVARDQD----LVQEDLHKLALWEQKWIMSFHPEKCITVHMSRSR----KQ
BfL2 -------------QIVLYADDTALF-YASKSVADINR---ALNADLCNIEKWLDANRLTLNVRKCKSMLFGTIRRRLRLET
BfRTE CN-PPTMYSKRVSCL-LYADDLVLFSESK----------QGLQCSLNRLEEYCQTWHLNVNLRKTKIIVFTKGGR---IP
BfNeSL QQHGYNLNDKYYITL-PFADDFCLITTNK----------RQHQKLITQISSNTKSMNLQLKPRKCKSMSIVSGK-----P
BfL1 AIHGLRLPGGREVKISQYADDNSAIVTTD----------ESIRRLFAVIDMYNRGSGSLLNLDKCMGLWLGSWR----KR
BfI -------------NFQYADDGAAW-KTGPNLRDLIT-HVELQ-IAKLISQWCPRWNMKVEESKTKALIFQANHQ--ADT
7 8
BfCR1 MEREYELHCHKLKACDQVKYLGITLTRDEKWSPHVTNITNK-ANRSFGFIRRNIKVN--SIAIREVAYKALVRPTLEYSS
BfL2 EELNLTLSGTYLEVVACFKYLGVWFDSCETWSIHINKLCST-VSSRLGVLRRLVPIL--PPKTLSMLFTCMILPKIEYCD
BfRTE KDVYFMYKNNPVEIVTNYCYLGIVVNSAGTFKANNKHLYSKGLKALFGINQSLDKADA-PLSVRNKLFDTCVKPIILYGS
BfNeSL SDVSFTIDGDPVKNTKDAPEEVLGWLHNLPE----------
BfL1 LDCPKNFKWTSVSIRLLGGTFGSVNMPLVNWRERQAKFEAV-LRRWDSFSLSFAGKVV-VVNNLALSTLWYIAPVFSPPE
BfI RVNHIRVNGKRIEVLPEIKLVGITLDDRETYTSHINNTKNKAYKALKVVSRVTNAKKNPNQEAHLLLYRALIRPILEYGA
9
BfCR1 SVWDLYTDKDIMTIEKVQRRAARWVCQRFRQTSSVGEMLESLQWETLQQRRKRARLITFFKIHHGIVTVNTSSPPTVKRQ
BfL2 IVWGNCGKSLSDNLQKLPNRAARLVLGLSHRSHVDNDHLSALGWNSLASRRKMHLLQTVFKSIHRQLPEYLQIFYRTSHT
BfRTE EIWGSVKSPKSCPIETIH----LKFCK---QTLHVPRSTSGLAARAELGRFPIHLEASLNAIKH-LIRLRQKVPADSFQS
BfNeSL ----------
BfL1 SVVKDITRAVFKFFWSSR------AEMVSRQTVMLPLDKGGWDLIDFGVKANCFYCRPFTNILD-----RPDLPWVQLAK
BfI ECTIMAGKKKEDAYAPIQ----------RRALEAATGCKTRTSTEALEALTGTIP----VDIHLTCRQAQA
BfCR1 TRLTRNVHPLTYVIPFCRTTYRQMSFFPRT-ILEWNSLPAETVTVPSIAAFREKMAHLN
BfL2 YATTLNSN-LSLQLPKVRLESGRBKFEYRG-AFSWNELPPAVKVASSALSFRK------
BfRTE DALSCQID-LDKAGAKCWASGVRQSLEECGYGYVW-HCPLQHNTNSSQIINSIN-----
BfNeSL ----------
BfL1 YWLGFFAR---RFEPSSWSNSSPHSPEAPPHYSLMRKLTLEFVNASASKTWDKICT---
BfI YLKLQQKK---TGQPHLRNHPT----------

**Table 1**. Non-LTR retrotransposons in protostomes and deuterostomes. The copy number for each clade, clade complexity (clades), total copy number (Copy num.), genomic burden (% Genome), and genome size (Gen. size) is shown.

| | *C. elegans* [12] | *D. melanogaster* [13] | *C. intestinalis* [7] and [14] | *B. floridae* | *T. rubripes* [15] | *R. norvegicus* [16] | *H. sapiens* [17] |
|---|---|---|---|---|---|---|---|
| CRE | | | | | 2,000 | | |
| I | | 67 | 9 | 3 | | | |
| Jockey | | 392 | | | | | |
| L1 | | | 22 | 32 | 500 | 597,000 | 904,000 |
| L2 | | | 24 | 35 | 6,500 | 48,000 | 408,000 |
| L3/CR1 | 1,000 | | | 25 | | 11,000 | 55,000 |
| LOA | | 18 | 69 | | | | |
| NeSL | 110 | | 6 | 6 | 30 | | |
| R1 | | 130 | | | | | |
| R2 | | 3-60 | 13 | | | | |
| R4 | | | | | 1,000 | | |
| Rex1 | | | | | 2,000 | | |
| RTE | 15 | | | 42 | 2,300 | | |
| Tad1 | | | | | | | |
| Clades | 3 | 5 | 6 | 6 | 7 | 3 | 3 |
| Copy num. | 1,115 | 667 | 143 | 143 | 14,300 | 657,000 | 1,368,000 |
| % Genome | <5 | <3 | <1 | <1 | 1.3 | 23.1 | 21.05 |
| Gen. size | $9.7 \cdot 10^7$ | $1.6 \cdot 10^8$ | $1.8 \cdot 10^8$ | $5.8 \cdot 10^8$ | $4 \cdot 10^8$ | $2.9 \cdot 10^9$ | $3.2 \cdot 10^9$ |

**Figure 2**. Schematic representation of the amphiouxus non-LTR elements and the phylogenetic relationships. (A) Phylogenetic tree based on the reverse transcriptase sequence with only the branch points (and neighbor-joining bootstrap support) leading to the major 14 clades of non-LTR elements. (B) Schematic representation of the characterised domains. The APE domains and RT blocks are numbered below each element. (C) The main features (length, copy number, assembled composites, range of similarity) defining each non-LTR retrotransposon clade are indicated




Not only the reported missing domains of the elements, but the elusive target repeats at the recipient site that would help to identify the borders at 5' and 3' of the elements, together with the fact that, in all genomes, most non-LTR copies are truncated at 5', strongly suggests that only a very reduced proportion of the identified retrotransposons are full-length copies with preserved autonomy. And those few, if any, could have remained undetectable in the raw genome database as, indeed, we have not found any full-length element.

The RT domain of non-LTR elements was used to establish the phylogenetic relationships of the amphioxus elements and the 14 reported non-LTR clades. In the neighbor-joining tree (Figure 2A). Twelve out of 14 clades were supported with

significant bootstrap values (≥70%). Clade I, showed the lowest bootstrap value (39%), in agreement with previous analyses [2, 10] whereas clade NeSL, gave bootstrap value close to the cut-off value (67%). Consequently, amphioxus composites were clustered in six different clades: CR1, I, L1, L2, NeSL and RTE (bootstraps: 99%, 39%, 82%, 94%, 67% and 71%, respectively) and were recorded as new members of each group.

## 4. Discussion

The approach used in this work has allowed the in silico identification and reconstruction of amphioxus non-LTR retrotransposons. The fact that the deduced features of *Bf*CR1, one of the derived elements, are in agreement with previous experimental findings [18] validates the in silico strategy and supports the data generated for other elements. We therefore propose that the *B. floridae* genome accommodates the old LINE NeSL, and young LINEs such as CR1, I, L1, L2 and RTE, with an overall structure consistent with that reported for each clade. Concerning *Bf*CR1, *Bf*L1, *Bf*L2 and *Bf*RTE retrotransposons, although no structural hallmarks for ORF1 and NABD could be confidently detected, the RT and APE domains were clearly ascertained. On the other hand, although the phylogenetic affiliation of *Bf*I and *Bf*NeSL was poorly supported, their ascription to the I and NeSL clades was established following the BLAST hits with reported retrotransposons (2e-46 for the I element of *Biomphalaria glabrata*, and 2e-33 for NeSL of *C. elegans*). The difficulties in *Bf*I and *Bf*NeSL characterization are probably due to their low copy number, 3 and 6 respectively, significantly lower than that of the other elements and because these clades are still weakly defined [6].

The estimation of the copy number of each element suffered from small inaccuracies caused by the cut-off e-value assigned to discriminate the sequences belonging to the same clade and, the fact that a whole genome shotgun sequencing does only yield a fraction of the genome. Nevertheless, the *in silico* estimates for *Bf*CR1 (25 copies) were in agreement with those obtained following an experimental approach (15) [17]. Our data showed low copy number per haploid genome for all the amphioxus elements, ranging from 3 to 42, a figure clearly similar to the number of the different composites assembled with the TIG clustering tools, thus showing the efficiency of the assembling procedure. Despite this overall scarcity, differences among clades were observed: *Bf*CR1, *Bf*L1, *Bf*L2 and *Bf*RTE were more frequent than *Bf*I and *Bf*NeSL elements. The permissiveness of the APE mediated insertion could account for the relative abundance of the former, whereas self regulatory mechanisms [19] or a high target site specificity [20] could explain the reduced number of the latter.

The mechanisms controlling copy number are still an open question but the values obtained in this work agree with those found in another lower chordate, *Ciona intestinalis*, and other organisms with small genomes such as *Drosophila melanogaster* and *Anopheles gambiae* [13, 21]. The overall copies of non-LTR retrotransposons in lower chordates represent, indeed, a very modest fraction of the genome, if compared to vertebrates (i.e. <1% in ascidians and amphiouxus versus the >20% in human). Then, low copy number in small genomes could easily be under self-control without having to invoke to host-promoted repression through methylation, as it has been shown in vertebrates and already discarded for *Bf*CR1 and *C. intestinalis* non-LTR retrotransposons [7]. Other mechanisms, such as co-supression for the I elements of *Drosophila* [19] or RNA silencing in fungi, plants and animals [22-24] could play a major role in the regulation of the expansion of this type of elements.

In summary, the present work shows that the amphioxus genome harbors at least 6 different clades of non-LTR retrotransposons, all present at low copy numbers. Although from our data we cannot assume that the overall structure of the amphioxus genome resembles that of the chordate *Ciona intestinalis*, it seems clear that both share a comparable burden of non-LTR retrotransposons. The analysis of the non-LTR content of the *B. floridae* genome here reported provides valuable data to understand the evolution of chordate genomes, enlarges the view of the distribution of the non-LTR clades in eukaryotes and highlights the structural differences between pre-vertebrate and vertebrate genomes.

## Acknoledgments

We thank JL Gelpí and M Pignatelli for help in the linux platform installation. This study was supported by the Ministerio de Educación y Ciencia (grant BMC 2003-05211). J.P. was the recipient of a fellowship from the Universitat de Barcelona.

## Conflict of interests

The authors have declared that no conflict of interests exists.

## References

1. Finnegan DJ. Eukaryotic transposable elements and genome evolution. Trends Genet 1989; 5:103-107.
2. Lovsin N, Gubensek F & Kordi D. Evolutionary dynamics in a novel L2 clade of non-LTR retrotransposons in Deuterostomia. Mol Biol Evol 2001; 18:2213-2224.
3. Holland LZ, Laudet V & Schubert M. The chordate amphioxus: an emerging model organism for developmental biology. Cell Mol Life Sci 2004; 61:2290-2308.
4. Altschul SF, Madden TL, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acid Res 1997; 25(17):3389-3402.
5. Swindell, SR & Plasterer TN. SEQMAN. Contig assembly. Methods Mol Biol 1997; 70:75-89.
6. Van Dellen K, Field J, et al. LINEs and SINE-like elements of the protist Entamoeba histolytica. Gene 2002; 297:229-239.
7. Permanyer J, Gonzalez-Duarte R & Albalat R. The non-LTR retrotransposons in Ciona intestinalis: new insights into the evolution of chordate genomes. Genome Biol 2003; 4:R73.

8. Thompson JD, Gibson TJ, et al. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 1997; 25:4876-4882.
9. Page RD. TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci 1996; 12:357-358.
10. Malik HS, Burke WD & Eickbush TH. The age and evolution of non-LTR retrotransposable elements. Mol Biol Evol 1999; 16:793-805.
11. Martin F, Olivares M, et al. Do non-long terminal repeat retrotransposons have nuclease activity? Trends Biochem Sci 1996; 21:283-285.
12. Zagrobelny M, Jeffares DC & Arctander P. Differences in non-LTR retrotransposons within C. elegans and C. briggsae genomes. Gene 2004; 330:61-66.
13. Berezikov E, Bucheton A & Busseau I. A search for reverse transcriptase-coding sequences reveals new non-LTR retrotransposons in the genome of Drosophila melanogaster. Genome Biol 2000; 1:RESEARCH0012.
14. Kojima KK and Fujiwara H. Cross-genome screening of novel sequence-specific non-LTR retrotransposons: various multicopy RNA genes and microsatellites are selected as targets. Mol Biol Evol 2004; 21(2):207-217
15. Aparicio S, Chapman J, et al. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 2002; 297:1301-1310.
16. Gibbs RA, Weinstock G, et al. Genome sequence of the brown norway rat yields insights into mammalian evolution. Nature 2004; 428:493-521.
17. Lander ES, Linton LM, et al. Initial sequencing and analysis of the human genome. Nature 2001; 409:860-921.
18. Albalat R, Permanyer J, et al. The first non-LTR retrotransposon characterised in the cephalochordate amphioxus, BfCR1, shows similarities to CR1-like elements. Cell Mol Life Sci 2003; 60:803-809.
19. Jensen S, Gassama MP & Heidmann T. Taming of transposable elements by homology-dependent gene silencing. Nat Genet 1999; 21:209-212.
20. Zingler N, Weichenrieder O, & Schumann GG. APE-type non-LTR retrotransposons: determinants involved in target site recognition. Cytogenet Genome Res 2005; 110:250-268.
21. Holt RA, Subramanian GM, et al. The genome sequence of the malaria mosquito Anopheles gambiae. Science 2002; 298:129-149.
22. Rossignol JL & Faugeron G. Gene inactivation triggered by recognition between DNA repeats. Experientia 1994; 50:307-317.
23. Waterhouse PM, Wang MB & Lough T. Gene silencing as an adaptive defence against viruses. Nature 2001; 411:834-842.
24. Bingham PM. Cosuppression comes to the animals. Cell 1997; 90:385-7.

**Capítol 3: Retrotransposons non-LTR a *Myxine glutinosa***

Article 4: *The non-LTR retroposon from* Myxine glutinosa*, MgLINE, is a frequent component of a basal vertebrate genome.*
Article per a ser sotmès a Cellular and Molecular Life Sciences

Els retrotransposons non-LTR representen una fracció molt important dels genomes seqüenciats dels vertebrats. En contraposició, aquests elements són quasi inexistents en els genomes dels cordats no vertebrats. Malauradament, els genomes dels vertebrats més primitius com els àgnats o vertebrats no mandíbulats romanen sense caracteritzar. Partint doncs d'aquesta mancança i donada la seva posició filogenètica a la base dels vertebrats, ens varem proposar identificar aquets tipus d'elements a l'àgnat *M.glutinosa*. En aquest treball es descriu MgLINE el primer retrotransposó non-LTR que pertany al clade dels elements CR1 i presenta les característiques típiques d'aquests elements. Amb una estima de 23000 còpies per genoma haploid, el número de còpies és moderadament elevat, clarament superior als observats a l'ascidi i l'amfiox però sensiblement inferior al dels elements més abundants dels vertebrats mandibulats. En quant la seva localització, sembla que almenys els representants que hem analitzat d'aquest grup es trobarien en la fracció metilada del genoma. Aquesta dada recolzaria el paper de la metilació com a senyal per a reprimir l'expressió d'aquests elements en els vertebrats primitius . La presència d'un element molt semblant a MgLINE en el genoma del mixinoideu *Eptatretus burgeri* suggereix que aquest tipus d'elements ha estat evolutivament conservat i possiblement es manté actiu en el llinatge dels àgnats.

# The non-LTR retroposon from *Myxine glutinosa*, MgLINE, is a frequent component of a basal vertebrate genome

**Permanyer J, Albalat R and Gonzàlez-Duarte R***

Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Av. Diagonal 645, 08028 Barcelona (Spain), Fax (+34)934110969, email: rgonzalez@ub.edu

**Abstract.** Non-LTR retrotransposons make up large portions of vertebrate genomes. These elements are diverse and widely dispersed among eukaryotes. They can be divided into at least 14 distinguishable clades, whose presence and abundance differs greatly beetwen chordate genomes. In lower chordates, like ascidians and amphioxus, there are few elements that belong to more than 5 clades. In jawed vertebrates, like humans, they could represent an important fraction of the genome with elements of few clades. In order to determine when the non-LTR retrotransposon explosion happened, we have examined the genome of the jawless vertebrate *Myxine glutinosa*. We have characterized MgLINE, a non-LTR retrotransposon that belongs to the CR1 clade. MgLINE is found in moderately high copy numbers in the mixyne genome which suggest that this kind of elements colonyzed early the vertebrate genomes, prior the agnatha-gnathostomata split.

**Keywords:** chordate, non-LTR retrotransposon, Genome evolution, mobile element

## Introduction

Jawless vertebrates, hagfishes and lampreys, are considered basal vertebrates, the former being at the most basal position (revised in [1]). Their genomes have not yet been well characterized but, it has been widely documented that lineage-specific duplications [2] contributed to expand the size of their genomes and provided the raw material for the non-LTR retrotransposon colonisation.

Retrotransposons are mobile genetic elements present in almost all eukaryotic genomes. They are rare in protostomes but frequent components of the genomes of some plants and jawed vertebrates (for instance [3] and [4]. All retrotransposons are distinguished by a life cycle that involves an RNA intermediate. Retrotransposons fall into two main categories depending on their terminal structures with or without long terminal repeats.

Non-LTR retrotransposons, classically divided in at least 14 clades [5], have been lately reorganised in 5 groups to cope with the increasing complexity of the cladistic classification [6]. These elements can additionally be grouped in two great clusters, *old* LINEs or site-specific endonuclease retrotransposons encoded in a single open reading frame (ORF), and *young* LINEs or non-site-specific endonuclease retrotrans-posons that encode two ORFs (ORF1 and ORF2) [7]. All of them, codify a reverse transcriptase (RT), the only common domain that is accompanied by additional motives. Not only RT but also an endonuclease activity is is required for their life cycle. The latter could be related to restriction enzymes (REL-endo), a feature of the *old* LINEs, or similar to an apurinic/apyramidinic endonuclease (APE) only found in *young* LINEs. Regardless of the type, transposable elements can no loguer be considered pure "junk". They participate in overall genome size, and serve as potent modulators of the evolution of the genome (revised in [8], [9]).

The chordate phylum is divided in 3 subphyla, urochordates, cephalochordates and vertebrates. Conventionally, cephalochordates had been considered the closest living relatives of the vertebrates although recent molecular data supported that this relationship applied instead to the urochordates [10]. The chordate phylum is characterized by an increase of the genome size due to subsequent total or partial genomic duplications in the vertebrate lineage [11]. Transposable ele-ments are usual components of the vertebrtae genomes. They have been well stablished for the jawed vertebrates (for instance in the human genome, [4], the urochordates [12], [13] and the cephalochordates [14], [15], [16] but this is not the case for the jawless vertebrates or agnathans, which despite their key phylogenetic position have been poorly studied. Indeed, the only elements reported are the mariner Tes1 from *Eptatretus stouti* [17] and the *old* LINE R2Eb from *Eptatretus burgeri* [18].

* Corresponding author

We have aimed to identify the non-LTR elements of M.glutinosa as key contributors of genome evolution at the base of the vertebrate origin. To that end, highly degenerated oligonucletotides were designed to amplify any non-LTR member and positive fragments were used to screen a genomic library constructed from a single specimen. The frequency and the low variation observed for the characterised non-LTR retrotransposon, MgLINE, allows to infer that the agnathan mobile genomic fraction is close to the "top" complexity adscribed to jawed vertebrates

## MATHERIALS AND METHODS

### Cloning and sequence of MgLINE

*Myxine glutinosa* animals were kindly provided by the Kristinebergs Marina Forskningsstation (Kristineberg, Sweeden). Total genomic DNA was isolated using the guanidine isothiocyanate method (Chirgwin 1979) with minor modifications. A single-animal *M. glutinosa* genomic library was constructed with Lambda FIX-II/*Xho*I partial fill-in vector (Stratagene).

With the alignment of 90 aminoacidic sequences of the non-LTR retroposon RT domain, 7 degenerated oligonucleotides were designed to amplify any non-LTR RT sequence (Figure 1A). PCR amplifications were performed with 100 to 500 pg (usually 400 pg) of genomic DNA and 1 U Taq DNA polymerase (BioTherm) in 25 µl of reaction volume containing 0.2 to 1 mM (usually 1 mM) for each oligonucleotide, 32mM each dNTP and 1 to 4 mM $MgCl_2$ (usually 2 mM). The general conditions were as follows: the initial denaturation step at 94°C for 5 min was followed by 40 cycles at 94°C for 30 sec, 50°C for 30 sec and 72°C for 1 min and a final extension step at 72°C for 5 min. Annealing temperatures ranging from 42ºC to 62ºC, extension step ranging from 30 sec to 150 sec and different PCR reagents combinations were tested.

The PCR products were cloned in a pUC 18 plasmid and sequenced using the Big Dye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) in a 3730 DNA Analyzer (Applied Biosystems).

An amplified fragment corresponding to the domains 0 to 4 [7] of a non-LTR RT was labeled with [α-$^{32}$P]dCTP by random-hexamer priming and used as probe to screen the genomic library. A low stringency hybridization was carried out in phosphate/SDS solution [19] at 55 °C overnight. Two 15-min washes were performed at 65°C in 2 × SSC, 0.1% SDS, 2 × 15 min at 65°C in 1 × SSC, 0.1% SDS and 1 × 15 min at 65°C in 0.2 × SSC, 0.1% SDS. Hybridization signals were detected by autoradiography. Only the signals present in the original and duplicated filters were considered. DNA fragments from positive recombinant phages were isolated through *Not*I digestion and cloned into a pBluescript KS II. The recombinant plasmids were characterized by restriction mapping. Random sequences were generated with the Genome Priming System (New England Biolabs) and analyzed and assembled with the Phred/Phrap/Consed Package (University of Washington).

### Copy number and genomic features

For quantitative slot-blot analyses, 200 ng, 100 ng, 50 ng and 25 ng of *Eco*RI-digested *M.glutinosa* genomic DNA and a serial dilutions of the plasmid-containing probe, *Eco*RI-restricted and mixed with 1 µg mouse genomic DNA (as nonspecific DNA) were denatured with 0.4 M NaOH and 25 mM EDTA in a final volume of 200 µl and blotted on Hybond-N nylon filters (Amersham Pharmacia Biotech) with a Minifold II slot-blot device (Schleicher & Schuell). Three genomic DNA replicates of three different isolated animals and three probe serial dilutions were performed. Before sample loading, the membrane was soaked in water and then neutralized with 2 M sodium acetate, pH 5.4 and fixed with UV light. Membranes were hybridized with the same probe used for library screening at the same hybridization and washing conditions. The slot-blot signal was quantified with the Personal Molecular Imager FX (Bio-Rad).

For Southern analyses 10 µg of *M. glutinosa* genomic DNA digested with *Eco*RI, *Hpa*II or *Msp*I was resolved on 0.7% agarose gels and transferred to nylon membranes. Southern blots were hybridized with the same probe used for library screening at identical hybridization and washing conditions.

### Phylogenetic analysis

The comparison of the sequenced phages allowed us to reconstruct a consensus sequence which was used to predict the protein corresponding to the non-LTR element. The RT deduced sequence was added to a previous alignment [7] where some sequences were removed in order to reduce the complexety, analysing elements of the Jockey group alone, and a new one was generated with Clustal X [20], maintaining the same pairwise gap penalties and multiple alignment parameters. Phylogenetic analyses were performed using the neighbor-joining method and drawn with the TreeViewPPC program

[21]. Confidence in each node was assessed by 1,000 bootstrap replicates.

## RESULTS AND DISCUSSION

### Identification of MgLINE, an element of the Jockey group in the *Myxine glutinosa* genome.

The analysis of vertebrate genomes have revealed the important fraction that correspond to repetitive DNA, mainly composed of transposable elements. Sometime ago, this DNA was considered as "junk DNA" [22] but the prevalence of the transposable elements in almost all genomes points to a beneficious role of these in the host genomes. In order to characterize the TE fraction of the *Myxine glutinosa* genome, a specific search of non-LTR elements through a library screening with a probe corresponding to a non-LTR RT was performed.

The highly degenerated oligonucleotides used in this work had been designed to amplify any non-LTR RT sequence. Notwithstanding the wide range of PCR conditions that have been tested, only the pairs 0F-4AR+4BR, 2F-5R and 4AF+4BF-5R yielded the expected RT product (Figure 1B). For each positive PCR, the fragments were cloned into a pUC18 and several clones were sequenced. Even some variation among different clones was observed, we were not capable to obtain sequences corresponding to non-LTR retroposons of different clades from the same PCR product. The inability to amplify RT sequences that belongs to elements from different clades indicates that the *Myxine glutinosa* genome could be similar to other vertebrate genomes where there is a main kind of non-LTR element like the human genome and L1 elements [4].

The PCR product 10.5, the one which covers more domains, was labeled and used to screen a *Myxine glutinosa* genomic library. The screening resulted in aproximately 38000 positive phages, 12 of them were isolated and characterized by restriction map analisys. Three of them with different patterns were completely sequenced. The characterized phages allowed us to reconstruct a consensus sequence of 2922 bp (Figure 1C and 2) that showed clear similarity with known non-LTR elements in blastx comparisons (6e-80 with *Bf*CR1 from *Branchiostoma floridae*, 7e-77 with SR1 from

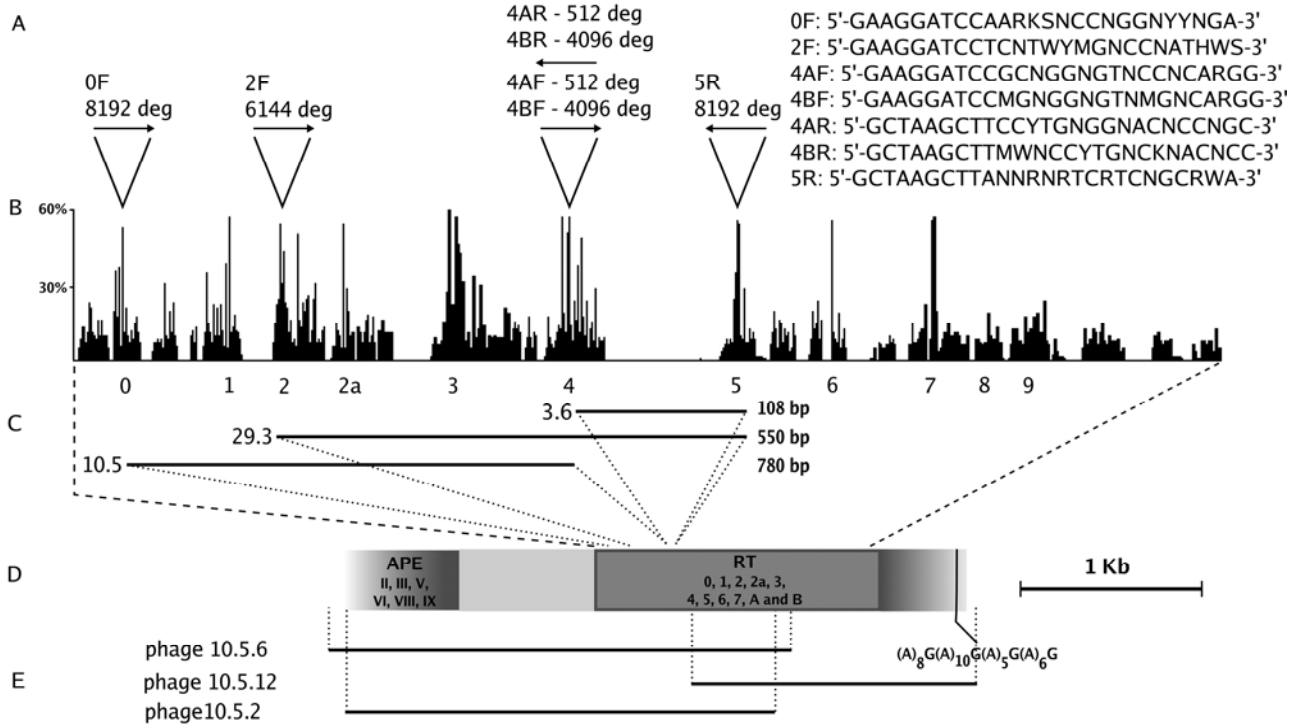



**Figure 1:** 7 degenerated oligonucleotides were designed (A) with the aim of a similarity plot (B) from the alignment of 90 protein RT sequences. Only some pairs amplified the expected RT product (C). The consensus MgLINE (D) generated from the comparison of 3 phages (E) showed the apurinic/apyramidinic endonuclease (APE), the reverse transcriptase (RT) and a putative terminal repeat. Note that the panels are in different scales.

*Schistosoma mansoni* and 3e-67 with CR1 from *Gallus gallus*, the best non-authomatic predicted hits) that belongs to the CR1 clade of the Jockey group [6]. These comparisons allowed us to deduce the protein sequence of MgLINE which covers at least the expected ORF2 APE (on the basis of the reported domains II, III, V, VI, VIII and IX, [23] and RT (with the structural hallmarks defined as blocks 0, 1, 2, 2a, 3, 4, 5, 6, 7, and boxes A and B, [7] (Figure 1C). The APE seems to be truncated as no domain I was recognized. Eventhough, the presence of conserved residues upstream domain I and the similarity with other elements, reinforces the validity of the deduced protein. Even no ORF1 signatures were detected, the blast homology with elements of the Jockey group and the presence and order of the domains typical of this group, reinforces the adscription of this repetitive sequence as a non-LTR element.

Furthermore, blastn comparisons with the GenBank sequences reveals the presence of the same type of element on the genome of the agnathan *Eptatretus burgeri*. The high degree of identity (83% nucleotidic identity over 2730bp and 58% protein identity) between MgLINE and EbLINE reveal that this type pf element originated before the myxinidae radiation and a possible recent activity in both phyla. An alternative less probable explanation, would be that the observed similarity between MgLINE and EbLINE is due to an horizontal transfer event although the geographical isolation of the species impairs the likelihood of this argument.

As no complete elements could be assembled, the characteristic target site duplication cannot be determined. Eventhough it is worth to mention the presence of a degenerated repeat 245 bp downstream of the predicted stop codon. By the other hand, it has been described for some elements from the Jockey clade that the terminal repeats could be based on polyA tracks (reviewed in [24]. So, it remains unclear which of these repeats, if any, participates in or are artifacts of the target-primed reverse transcription (reviewed in [6] or represents a nearby repetitive genomic sequence not related with MgLINE.

**Phylogenetic relathionships**

The analysis of the 2922 bp consensus of the MgLINE revelead that both the blast matches, with 30% aminoacidic identity with BfCR1, and the domains order corresponds to elements of the Jockey group. In order to determine the clade at which the element MgLINE belongs, phylogenetic analysis based on the prdicted RT sequence were performed using the neighbour-joining method with sequences of elements of the Jockey group. The tree was rooted using *Neurospora* organellar group II intron (accession number S07649), considered the ancestor of all non-LTR elements. The phylogenetic analyses show an evident relationship with the CR1 elements of the Jockey group (data not shown). A second tree with elements of the CR1 clade and *Neurospora* organellar group II intron and Rex1 elements as outgroup was done (Figure 3). Although high bootstrap values were obtained for the CR1 clade (95%), an anomalous topology was observed in this clade. The tree presented an internal tetracothomy which clearly divides the CR1 elements in 4 groups, those belonging to *A. gambiae*, *C. elegans*, the group defined by MgLINE and the group defined by the canonical CR1 elements from *G. gallus*. The complicate arrangement of CR1 elements, with a possible horizontal transfer of the SR1 element, show a complex evolution of this kind of elements and reinforces the artificial classification of the non-LTR retroposons in clades which may be replaced to the classification in groups as each clade seems to accomodate elements with different evolutionary histories [25].

**Copy number and genomic features of MgLINE**

Copy number of MgLINE was determined through two independent experimental approaches –slot blot and genomic library screenings–. We performed a quantitative slot blot in order to compare the degree of hybridisation between serial dilution of genomic DNA from three different individuals and serial dilutions of a probe-containing plasmid. This approach showed a copy number ranging from 5,000 to 23,000 copies per haploid genome depending on the individual. In addition to this method, a library screening on 7 200,000-phage plates (high density plates with ~500 phages/$cm^2$) and obtained an average of 5,400 positives per plate. Considering an average insert size 15 Kb, and a hagfish haploid genome of $4.25x10^9$ bp, this resulted in 7,000 copies per haploid genome. As the results were slightly unequal between the two approaches, we redid the genomic library screening hybridising 9 low denisity phage plates (15 phages/$cm^2$) to obtain a more accurated estimate. In this case, we obtained

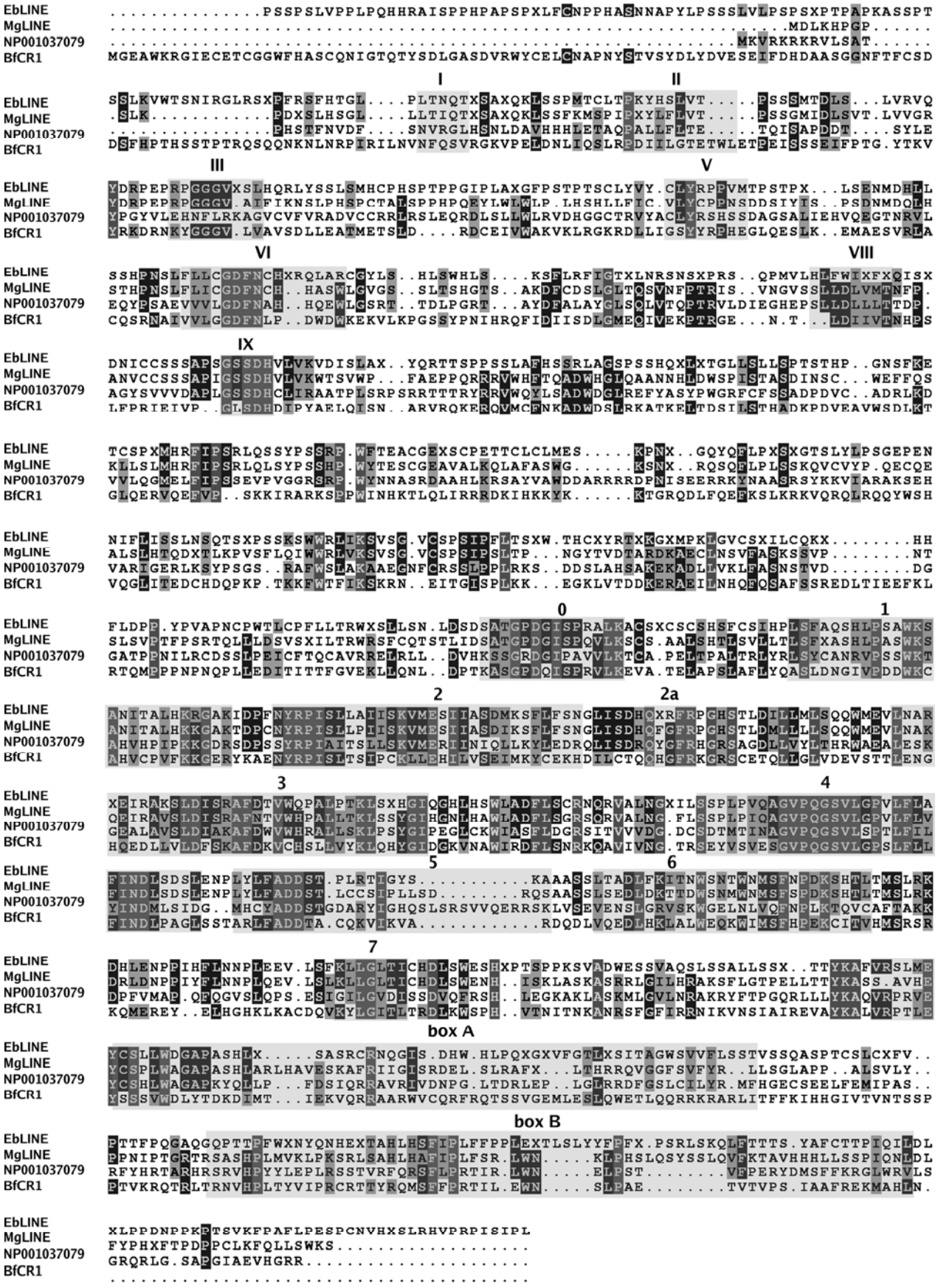



**Figure 2:** Alignment of deduced protein sequence of CR1 clade members. N-terminal endonuclease domains I, II, III, V, VI, VIII and IX, 0-7 conserved blocks found in all RTs and boxes A and B in the C-terminal region are indicated. Amino acid conservation and similarity are shown in black and grey backgrounds respectively. EbLINE corresponds to the deduced non-LTR retroposon from *E. burgeri* (accession number AY965678), MgLINE to the element described in this work, NP001037079 to an authomatic predicted element from *B. mori* and BfCR1 to the CR1 element described in *B. floridae* [14].

1,040 positive phages over 6,012 screened phages which implied 52,000 copies per haploid genome. The observed differences between the two copy number estimations by library screening are due to the high density of positive phages observed in the hybridization of the 7 200,000-phage plates where more than one positive phage is in the same autoradiography dot signal. An estimation of the fraction of the genome harboring MgLINE was done assuming that the genome accomodates 52,000 full-length copies with an average size of 5 Kb, the value obtained would represent less than 6% of the haploid genome. Furthermore, the presence of at least 2 different elements in the genome of *E. burgeri*, the EbLINE described in this work and the element EbR2 [18], suggest that the *M. glutinosa* genome could harbour more than one kind of element which would increase the mobile burden.

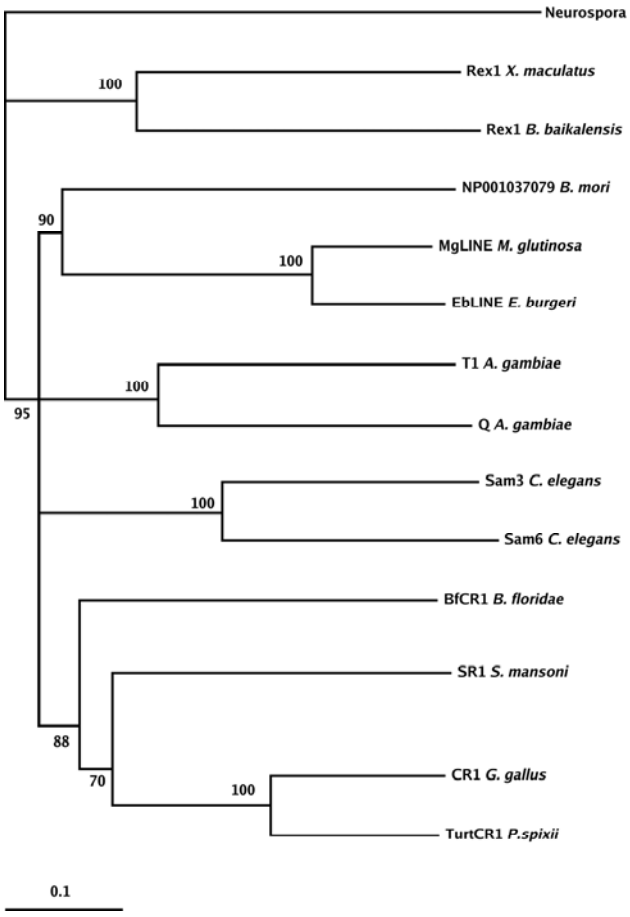



**Figure 3:** Neighbour-joining phylogenetic tree from the alignment of the RT domain of CR1 retroposons. Nodes are assessed from 1000 bootstrap replicates. The tree is rooted using the Neurospora organellar group II intron, considered as the ancestor of all retroposons, and 2 Rex1 elements as a sister group of the CR1 elements.

This implies an important increase of copies compared to the pre-vertebrates counterparts, less than 200 non-LTR elements in the *Ciona intestinalis* [13] and *Branchiostoma floridae* genome [15], and in the same range of some actinopterygian fishes, 14,300 non-LTR elements in *Takifugu rubripes*, 2896 in *Tetraodon nigroviridis* and more than 4500 in *Danio* rerio [5]. This increase in copy number could be related to the genome expansion proposed for the vertebrata phyla [11] which provided the raw material to buffer the TE-induced genomic rearrangements. The methylation status of the genomic enviroment of MgLINE was determined by comparing the hybridisation patterns of genomic DNA restricted with the methylation-sensitive enzyme *Hpa*II, and the methylation-insensitive isoschizomer *Msp*I. The different *Hpa*II and *Msp*I patterns supports the methylated genomic enviroment of this element (data not shown) which was expected as agnathans have extensively methylated genome [26] and could help to understand the high numbers observed. Although the role for DNA methylation in TEs control is still controversial [27] because seems not to apply in the small moderately methylated genomes of the basal chordates [13]. Seems that, in the vertebrate genomes, transposable elements accumulate in large numbers not despite, but precisely because of, the mechanisms that make large, redundant genomes possible [6].

## CONCLUSSIONS

The *Mixine glutinosa* genome harbours at least one kind of non-LTR retroposon, named MgLINE, that belongs to the CR1 clade of the Jockey group and presents the expected domains of this kind of element. Although no full length elements were detected, the high copy numbers observed and the high identity within MgLINE and EbLINE suggest the presence of active elements and so, full length. As in vertebrates, MgLINE is present in the methylated fraction of the genome supporting the theory that the mechanisms to control gene expression allowed the important raise in TE copy numbers observed in the vertebrates.

*Acknowledgments.* We wish to thank the Serveis Científico-Tècnics at PCB. This study was supported by the Ministerio de Educación y Ciencia (grant BMC 2003-05211). J.P. was recipient of a UB fellowship

1.Halanych, K.M., *The new view of animal phylogeny.* Annual Review of Ecology, Evolution, and Systematics, 2004. **35**(1): p. 229-256.
2.Lundin, L.G., *Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse.* Genomics, 1993. **16**(1): p. 1-19.

3.Vicient, C.M., et al., *Retrotransposon BARE-1 and Its Role in Genome Evolution in the Genus Hordeum.* Plant Cell, 1999. **11**(9): p. 1769-1784.
4.Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
5.Volff, J.N., et al., *Diversity of retrotransposable elements in compact pufferfish genomes.* Trends Genet, 2003. **19**(12): p. 674-8.
6.Craig, N.L., *Mobile DNA II*. 2001, Washington, D.C.: ASM Press.
7.Malik, H.S., W.D. Burke, and T.H. Eickbush, *The age and evolution of non-LTR retrotransposable elements.* Mol Biol Evol, 1999. **16**(6): p. 793-805.
8.Finnegan, D.J., *Eukaryotic transposable elements and genome evolution.* Trends Genet, 1989. **5**(4): p. 103-7.
9.Makalowski, W., *Genomic scrap yard: how genomes utilize all that junk.* Gene, 2000. **259**(1-2): p. 61-7.
10.Delsuc, F., et al., *Tunicates and not cephalochordates are the closest living relatives of vertebrates.* Nature, 2006. **439**(7079): p. 965-8.
11.Ohno, S., *Evolution by gene duplication*. 1970, Springer-Verlag: Heidelberg.
12.Simmen, M.W., et al., *Nonmethylated transposable elements and methylated genes in a chordate genome.* Science, 1999. **283**(5405): p. 1164-7.
13.Permanyer, J., R. Gonzalez-Duarte, and R. Albalat, *The non-LTR retrotransposons in Ciona intestinalis: new insights into the evolution of chordate genomes.* Genome Biol, 2003. **4**(11): p. R73.
14.Albalat, R., et al., *The first non-LTR retrotransposon characterised in the cephalochordate amphioxus, BfCR1, shows similarities to CR1-like elements.* Cell Mol Life Sci, 2003. **60**(4): p. 803-9.
15.Permanyer, J., R. Albalat, and R. Gonzalez-Duarte, *Getting closer to a pre-vertebrate genome: the non-LTR retrotransposons of Branchiostoma floridae.* Int J Biol Sci, 2006. **2**(2): p. 48-53.
16.Osborne, P.W., et al., *Identification and characterisation of five novel miniature inverted-repeat transposable elements (MITEs) in amphioxus (Branchiostoma floridae).* Int J Biol Sci, 2006. **2**(2): p. 54-60.
17.Heierhorst, J., K. Lederis, and D. Richter, *Presence of a member of the Tc1-like transposon family from nematodes and Drosophila within the vasotocin gene of a primitive vertebrate, the Pacific hagfish Eptatretus stouti.* Proc Natl Acad Sci U S A, 1992. **89**(15): p. 6798-802.
18.Kojima, K.K. and H. Fujiwara, *Long-term inheritance of the 28S rDNA-specific retrotransposon R2.* Mol Biol Evol, 2005. **22**(11): p. 2157-65.
19.Church, G.M. and W. Gilbert, *Genomic sequencing.* Proc Natl Acad Sci U S A, 1984. **81**(7)): p. 1991-5.
20.Thompson, J.D., et al., *The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.* Nucleic Acids Res, 1997. **25**(24)): p. 4876-82.
21.Page, R.D., *TreeView: an application to display phylogenetic trees on personal computers.* Comput Appl Biosci, 1996. **12**(4)): p. 357-8.
22.Ohno, S., *So much "junk" DNA in our genome*. Brookhaven Symposia in Biology. 1972.
23.Martin, F., et al., *Do non-long terminal repeat retrotransposons have nuclease activity?* Trends Biochem Sci, 1996. **21**(8): p. 283-5.
24.Zingler, N., O. Weichenrieder, and G.G. Schumann, *APE-type non-LTR retrotransposons: determinants involved in target site recognition.* Cytogenet Genome Res, 2005. **110**(1-4): p. 250-68.
25.Tu, Z., J. Isoe, and J.A. Guzova, *Structural, genomic, and phylogenetic analysis of Lian, a novel family of non-LTR retrotransposons in the yellow fever mosquito, Aedes aegypti.* Mol Biol Evol, 1998. **15**(7): p. 837-53.
26.Tweedie, S., et al., *Methylation of genomes and genes at the invertebrate-vertebrate boundary.* Mol Cell Biol, 1997. **17**(3)): p. 1469-75.
27.Bird, A., *Does DNA methylation control transposition of selfish elements in the germline?* Trends Genet, 1997. **13**(12): p. 469-72.