ISSN 1993-6842 (on-line); ISSN 0233-7657 (print)
Biopolymers and Cell. 2020. Vol. 36. N 1. P 14–22
doi: http://dx.doi.org/10.7124/bc.000A1D

UDC 577.21

Extreme diversity of SINE families in amphioxus *Branchiostoma belcheri*

S. A. Kosushkin, N. S. Vassetzky

Engelhardt Institute of Molecular Biology, Russian Academy of Sciences
32, Vavilova Str., Moscow, Russian Federation, 119991
nvas@eimb.ru

Short INterspersed Elements (SINEs) are an important component of the genome in higher eukaryotes. Commonly, their genomes have one or several SINE families. **Aim.** Identification and analysis of SINEs in the oriental lancelet *Branchistoma belcheri* was the goal of this study. **Methods.** Original and conventional bioinformatics methods were used. **Results.** Eighteen *B. belcheri* SINE families have been found and analyzed. **Conclusions.** Our analysis of the genome of *B. belcheri* has demonstrated an extreme variation of SINEs. The multitude of retrotransposons in lancelets is discussed in the context of chordate genomes.

K e y w o r d s: Short INterspersed Elements, *Branchistoma belcheri*, retrotransposons.

Introduction

Lancelets (Cephalochordata) are small jawless segmented marine filter-feeders with a notochord. Fossil data dates them to the early Cambrian (~530 mya). Lancelets can provide an insight into the evolution and development of vertebrates, and sequencing of some of their genomes is an exciting advancement in the genomics of vertebrates. *Branchiostoma belcheri* Gray is one of the few remaining species of lancelets found in South Africa, Indo-West Pacific, Madagascar, and China. Its genome has been sequenced recently [1].

The genomes are invaded by various repetitive elements, the most abundant of which (at least in higher eukaryotes) are Long (LINEs) and Short INterspersed Elements (SINEs). Both are retrotransposons, i.e., their amplification cycle includes the transcription of their genomic copies, reverse transcription of these transcripts, and integration of DNA into the genome. LINEs rely on the transcription by the cellular RNA polymerase II, while the reverse transcription and integration are fulfilled by their own enzymes. SINEs encode no enzymes and employ the cell machinery for their transcription (unlike LINEs, SINEs are transcribed by RNA polymerase III (pol III) similarly to tRNAs) and their partner LINEs

for the reverse transcription/integration. Accordingly, SINEs have pol III promoters for transcription and sequences recognized by the enzymes of their partner LINE for reverse transcription/integration.

A typical SINE consists of the *head* derived from one of cellular RNA species (tRNA, 7SL RNA, or 5S rRNA); the *body* the terminal part of which is recognized by the partner reverse transcriptase (RT); and the *tail*, a stretch of simple repeats. There are variations; certain SINEs have no body or their body contains sequences of unknown origin and function (some of them called *central domains*) are shared between otherwise unrelated SINE families, *etc.* [2]

LINEs are found in the genomes of all higher eukaryotes. Clearly, SINEs cannot exist without LINEs; however, there are rare species whose genomes have LINEs but lack SINEs (e.g., *Saccharomyces* or *Drosophila*). During evolution, LINE (sub)families become inactive; clearly, partner SINEs also cease to amplify. If another LINE family becomes active in a particular genome, replacing the sequence recognized by its RT can reanimate a SINE [3]. A high number of SINE genomic copies favors such a module exchange. Usually, a genome harbors one or a few SINE families; some of them are inactive and were amplified in the ancestors.

Okada and colleagues have discovered the first SINE in lancelets, BflSINE1 (Bfl1) in *Branchiostoma floridae* (Florida lancelet) [4]. Another one, BflSINE2 (Bfl2), can be found in the last Repbase Update (RU) [5] available to us (as a consensus sequence with no description). Both are tRNA-derived SINEs with the *CORE* or *Deu* central domains, respectively. Their LINE partner remained unknown. Canestro & Albalat also mentioned a single SINE family (Bfl1) in *B. floridae* [6]. Finally, an "*Alu*" element reported in *B. floridae* [7] is neither *Alu* nor SINE but rather a fragment of an RTEX LINE. The paper on the *B. floridae* genome sequencing only mentions that identified transposable elements were deposited in RU [8].

In contrast to SINEs, a variety of LINE families were reported in the Florida lancelet; these include CR1/L2, Crack, I/Vingi/Outcast, L1/Tx1, NeSL/Hero, Proto2, REX1, and RTE/RTEX [6].

As concerns *B. belcheri*, Huang *et al.* presented an in-depth study of lancelet genomes, which included transposon analysis without their detailed classification. However, nonautonomous transposons (SINEs) were not considered at all [1]. Apart from that, we have found only one publication on *B. belcheri* repetitive sequences but it mentions only microsatellites [9].

Materials and Methods

The *Branchiostoma belcheri* draft genome (assembly v18h27.r3, internal ID: bbv18) was downloaded from *The Lancelet (Branchiostoma belcheri) Genome Sequencing and Annotation Project Database* [10].

We used custom Perl scripts based on the Smith-Waterman search to find genomic copies of SINEs and LINEs with the desired similarity to family consensuses. After most SINE families were identified, the genome bank was successively depleted using their consensus sequences (from long to short). SINE families proved to share extended similar regions and the similarity between indi-

vidual copies could be low (the mean similarity varied from 63 to 95 %), which greatly complicated their identification. We had to adjust search parameters and use unique search patterns to avoid misidentifications. Overall, 13,715 full-length SINE sequences have been found.

Multiple alignments were generated and edited using MAFFT [11] and GeneDoc [12]; the mean similarity was determined for 100 randomly selected sequences (or all available if less) using the *alistat* program [13]. Other sequence analysis tools were used such as *fuzznuc*, *dotmatcher* [14], *etc.*

Search for LINE partners of SINEs was performed using the 3'-end sequence (usually <100 bp) of a SINE family consensus as a query against RU followed by a search of matching LINE consensus sequence(s) in *B. belcheri* genome.

All *B. belcheri* SINE families were deposited to SINEBase [2].

Results and Discussion

SINEs are composed of modules such as tRNA-derived region, central domain, LINE-derived region, and tail (simple repeat region). For clarity, we consider SINE family as a set of SINEs of a common origin composed of the same modules in the same order (with the exception of the tail, which can vary) [2]. Minor variations (deletions not spanning over the whole module, insertions, and substitutions) give rise to subfamilies.

B. belcheri SINE families

In addition to BflSINE1 and BflSINE2 (hereafter referred to as Bfl1 and Bfl2, respectively) described previously [4,5], we have identified 16 new SINE families in the genome of *B. belcheri*. The length of the consensus sequences ranges from 194 to 446 bp, and the number of their full-length genomic copies varies from 20 to ~6000. The mean similarity of the genomic copies (which assumingly corresponds to the family age) ranges from 63 to 95 %. There were less abundant variants but these were ignored (20 copies roughly correspond to 100 copies in mammals adjusted for the genome size).

All identified families are typical tRNA-derived SINEs. They can be divided into two pools of SINE families excluding the standalone Bfl2. The pool A is composed of the tRNA head followed by the central domains CORE and Deu; the pool B is similar but lacks the Deu domain. The central domains are followed by the LINE-derived sequences and/or sequences of unknown origin. Their 3'-end is the tail, a few nucleotides tandemly repeated several times (Figs 1 and 2).

TRNA-derived head. It is not always possible to identify the cellular tRNA that gave rise to a SINE since the sequences of certain tRNA species are quite similar, and the corresponding region in SINEs has more or less modifications. However, the head sequence similarity allows the identified SINE families to be divided into three groups (indicated by shades of green in Fig. 1). The first group includes 15 families; the second, BbeA8 and BbeB3; and the third, only BbeA7. Amazingly, the latter has an extra tRNA-derived region in reverse orientation downstream of the central domain.

We specifically searched for 5S rRNA- and 7SL RNA-derived SINE sequences but failed to detect any in the genome of *B. belcheri*.

Central domains. All *B. belcheri* SINE families have central domains, CORE, Deu, or both. Ten families have the CORE domain followed by the Deu (pool A), seven families have CORE domain only (pool B), and Bfl2 SINE has the Deu domain only. The similarity region largely covers the greater part of the CORE consensus sequence (orange line in Fig. 1); however, some are truncated (BbeB5 and BbeB2). Most Deu domains in *B. belcheri* have the same deletion relative to the consensus sequence (yellow line in Fig. 1) but are 3'-truncated to different extents, and they are quite similar to each other. Bfl2 has a dissimilar matching pattern with Deu (Fig. 1). Finally, BbeB6 has a relatively short Deu-related region in reverse orientation (yellow arrows in Fig. 1).

LINE-derived regions. The identification of *B. belcheri* LINEs was not our goal, we only wanted to identify LINE partners of SINEs. Accordingly, we searched for LINE sequences with matching 3' ends in the genome (the least significant was 67 % identity over 32 nt). It should be noted that most genomic copies of LINEs are not full-length. There are numerous LINE fragments of different lengths; apart from that they can have extended deletions or modifications relative to the full-length consensus sequences. An extreme variation of *B.*

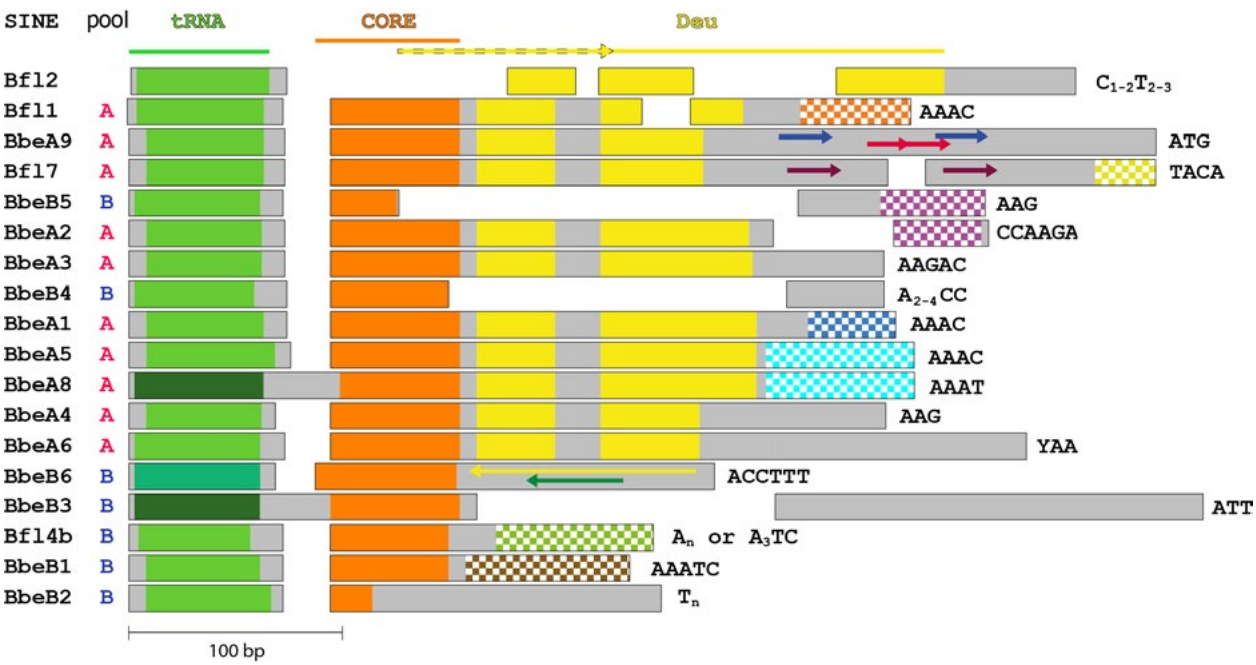

Fig. 1. Schematic alignment of *B. belcheri* SINE families. Pools (A or B) are indicated after the SINE name (red and blue, respectively). TRNA-related regions are shown in green, and different shades correspond to different tRNA species. Central domains CORE and Deu are indicated as orange and yellow lines above, respectively; and the corresponding SINE regions are in the same colors. LINE-derived regions are marked by checkerboard pattern, and the same colors indicate the same LINE partners. Different direct repeats are indicated by colored arrows. Tail repeat units are shown on the right. Dissimilar regions of unknown origin are shaded in gray.

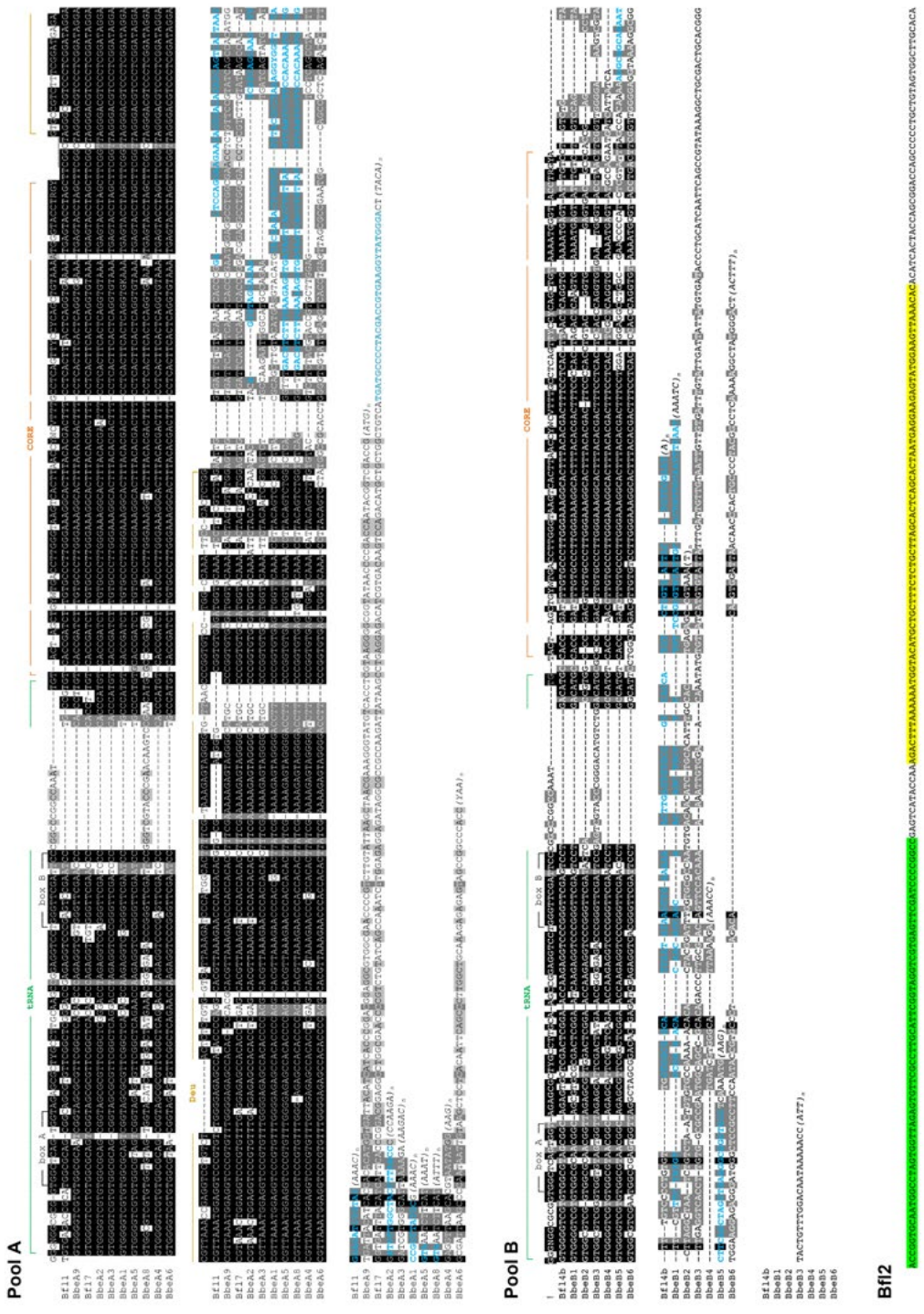

Fig.2. Consensus sequences of *B. belcheri* SINE families. Multiple sequence alignments for pools A (upper panel) and B (middle panel). The upper sequence (without a name) includes consensus sequences of SINE tRNA-derived region and CORE and/or Deu domains (which are indicated above together with pol III promoters). LINE-derived regions are indicated in blue font. The tail simple repeats are given in parentheses. The lower panel presents the Bfl2 consensus sequence with tRNA and Deu regions marked in green and yellow, respectively.

belcheri retrotransposons further complicates their analysis. However, we have identified the following LINE-partners of *B. belcheri* SINEs (Table).

Bfl1's LINE-partner belongs to the Crack clade, it is most similar to Crack-16_BF in RU; this SINE-LINE pair also shares the tail sequence $(AAAC)_n$. We have found two full-length copies of this LINE (84 % over 59 nt).

Bfl7 3' end terminus is similar to three genomic full-length sequences of an RTE clade LINE, RTE-18_BF and RTE-5_BF in RU (72 % over 32 nt).

All other SINE partners identified in *B. belcheri* genome belong to CR1 clade.

BbeA1. Thirty-six genomic sequences are similar to the 3' end of 36 copies of a LINE family most similar to CR1-1_BF in RU (85 % over 47 bp).

BbeA2 and BbeB5 have similar 3'-terminal sequences (82 % over 38 nt) suggesting that they share the LINE partner. BbeA2 is 81 % similar over 39 bp with two full-length genomic sequences of a CR1 clade LINE (CR1-7_BF in RU). The similarity of the region in BbeB5 with *B. belcheri* LINEs is lower (65 % over 39 nt) but more significant (89 % over 56 nt) with seven full-length *B. floridae* LINEs (CR1-7_BF in RU).

BbeA5 and BbeA8 have nearly identical 3'-terminal sequences (98 % over 77 bp). The genome of *B. belcheri* has 16 full-length copies of their LINE-partner, which can be divided into at least four different groups by the 3' end sequences. The most similar RU consensus sequences are CR1-14_BF, CR1-16_BF, CR1-27_BF, and CR1-47_BF (73 % over 79 nt for BbeA5 and 73 % over 74 nt for BbeA8).

The 3'-terminal regions of two SINE families (Bfl4b and BbeB1) each correspond to two different LINE partners; and all of them belong to the CR1 clade. One LINE partner of Bfl4b is CR1-28_BF (4 full-length copies, 95 % over 52 nt); the other corresponds to CR1-17_BF, CR1-51_BF, and CR1-54_BF in RU (40 full-length copies, 70 % over 90 nt).

BbeB1 partners are represented in RU by CR1-24_BF (5 full-length copies, 83 % over 85 nt) and CR1-33_BF, CR1-37_BF, and CR1-23_BF (14 full-length copies, 72 % over 82 nt).

Tails. The 3'-terminal simple repeats vary between SINE families. Actually, they could vary even within families (which is not unusual in SINEs). Some but not all SINE families have the same tail repeat sequence with their LINE partners: Bfl1, BbeA5, BbeA8, BbeB1 (with CR1-24_BF), and BbeB5 (underlined in Table).

Other features. Two SINE families (BbeA9 and Bfl7) have direct repeats (we searched for >70 % over >20 nt) in the body region downstream of the Deu domain (indicated by different colored arrows in Fig. 1). In addition, BbeA7 has an additional tRNA-related region in reverse orientation downstream of the CORE domain.

SINEs in Lancelets

The sequencing of lancelet genomes was an exciting scientific advancement, and the high variation of the genomic repetitive elements was noted. Unexpectedly, SINEs, which constitute a substantial volume in many eukaryotic genomes [15], remained unnoticed. Here we try to fill this gap.

The genome of *B. belcheri* harbors 18 SINE families, which amount to almost 14,000 co-

pies and ~1 % of the genome. All of them are tRNA-derived, although, from three different tRNA species. All families have central domains, CORE and/or Deu; accordingly, they can be conventionally divided into three groups: pool A (10 families) with CORE followed by Deu, pool B (7 families) with CORE only, and Bfl2 with Deu only. Notice that the Deu pattern is similar in all pool A SINEs (at least in the 5' part) but distinct from that in Bfl2 (Fig. 1).

A half of SINE families proved to share the 3'-terminal region with one of *B. belcheri* LINEs, which suggests that these LINEs mobilize them. Most LINE partners belong to the CR1 clade with two exceptions, RTE and Crack (for Bfl1 and Bfl7 SINEs, both of which are found in *B. floridae*). Two SINE pairs share similar 3'-end sequences, and thus, the same LINE partners, BbeB5/BbeA1 and BbeA5/BbeA8.

We can think of two reasons why LINE partners of nine SINE families have not been identified. (1) These LINEs do not require the 3' sequence to initiate the reverse transcription [16] or (2) they were active long ago and their sequences have degraded since then. The latter seems unlikely at least for the partner of BbeA4 (its mean sequence similarity is the highest of all *B. belcheri* SINE families suggesting the recent activity of both this SINE and its LINE partner). In addition, we could identify few full-length LINE partner copies due to sequencing errors not uncommon for repetitive sequences; or *B. belcheri* LINEs missing in RU could be omitted.

The number of full-length copies varies ~300-fold between families, from more than 6,000 to 20. Three most abundant families (Bfl4b, Bfl2, and BbeB1) amount to 65 % of the total number of SINE copies.

The mean similarity within a retrotransposon family can be considered a marker of their age. Thus, Bbe6 and BbeA7 (<70 %) were likely active before Bfl7, BbeA4, and BbeA8 (>90 %). Unexpectedly, no correlation was found between the mean similarities between SINE-LINE partners. This can be attributed to low numbers of full-length LINE copies and to their numerous subfamilies (data not shown). Overall, younger SINEs (>90 %) are few in number (20-155 copies) while two most abundant families are old (71-72 %). No such trend was observed in LINEs.

SINEs readily exchange their modules to maintain their amplification competence, which is mediated by both DNA-based mechanisms (common for all genomic elements) and RNA-related ones (specific for retrotransposons) [3]. The most explicable case is the replacement of LINE-derived regions when a new LINE variety becomes active. For instance, Bfl4b's LINE partner seems to be inactive (62 % mean similarity), and the corresponding region in BbeB1 was replaced with the 3'-terminal region of another active LINE (98 %), while all other regions of Bfl4b and BbeB1 remained largely unaltered (Fig. 1).

The case of BbeA5 and BbeA8 can also be explained. They differ only by the tRNA-derived head, and this replacement could be either positive or neutral (but not negative; BbeA8 shares the tRNA-derived region with BbeB3, which has more genomic copies than BbeA5 and BbeA8) (Fig. 1). The function of SINE central domains remains enigmatic and we will refrain from speculating about their rearrangements. This also applies to the tan-

dem repeats found in Bfl7 and BbeA9 as well as to the tRNA-derived region in reverse orientation found in BbeA7 (although such regions were found in a few mosquito and apple SINEs, [2]).

Surprisingly, no 5S rRNA-derived SINEs have been found in *B. belcheri*, although such SINEs exist in *B. floridae* (data not shown, one of them was mentioned in the Supplementary Materials of [8]).

The majority of genomes in higher eukaryotes contain SINEs, most commonly, one to three families; however, rare species can have six (wallaby *Macropus eugenii* and mosquito *Culex pipiens*) or even seven ones (sea urchin *Strongylocentrotus purpuratus*) [2]. The multitude of *B. belcheri* SINE families is exceptional, eighteen (the more so considering that lancelet genomes are substantially smaller compared to most other chordates). This reproduces the multitude of LINE families in lancelets. The reverse transcription machinery of LINEs is required for the amplification of SINEs, and their ability to exchange modules allows them to utilize different LINE partners. For instance, Bfl4b acquired another 3'-terminal region, which allowed the nascent BbeB1 to use a different LINE partner (Fig. 1).

It is hard to explain the revealed extraordinary multitude of SINEs and transposons in general since we are only breaking new grounds in cephalochordate genomics. However, *B. belcheri* transposons are not hypermethylated [1], which is also true for tunicates [17] but not for higher chordates (vertebrates). At the same time, the rate of lancelet evolution is similar to that in tetrapods [1]. As a matter of speculation, we can propose that the system of repression of transposable elements has not been established in lancelets, and their amplification was limited by negative selection against deleterious transposon insertions. New genomic data on lancelets and possibly tunicates can bring light to this problem.

Acknowledgments

We thank Dr. Dmitri Kramerov for critical reading of the MS.

Funding

This work was supported by the Russian Foundation for Basic Research (project no. 17-04-01723).

REFERENCES

1. *Huang S, Chen Z, Yan X, Yu T, Huang G, Yan Q, Pontarotti PA, Zhao H, Li J, Yang P, Wang R, Li R, Tao X, Deng T, Wang Y, Li G, Zhang Q, Zhou S, You L, Yuan S, Fu Y, Wu F, Dong M, Chen S, Xu A.* Decelerated genome evolution in modern vertebrates revealed by analysis of multiple lancelet genomes. *Nat Commun.* 2014;**5**:5896.
2. *Vassetzky NS, Kramerov DA.* SINEBase: a database and tool for SINE analysis. *Nucleic Acids Res.* 2013;**41**(Database issue):D83–9.
3. *Kramerov DA, Vassetzky NS.* Origin and evolution of SINEs in eukaryotic genomes. *Heredity (Edinb).* 2011;**107**(6):487–95.
4. *Nishihara H, Smit AF, Okada N.* Functional noncoding sequences derived from SINEs in the mammalian genome. *Genome Res.* 2006;**16**(7):864–74.
5. *Bao W, Kojima KK, Kohany O.* Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA.* 2015;**6**:11.
6. *Cañestro C, Albalat R.* Transposon diversity is higher in amphioxus than in vertebrates: functional and evolutionary inferences. *Brief Funct Genomics.* 2012;**11**(2):131–41.
7. *Holland LZ.* A SINE in the genome of the cephalochordate amphioxus is an Alu element. *Int J Biol Sci.* 2006;**2**(2):61–5.

8. *Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu JK, Benito-Gutiérrez EL, Dubchak I, Garcia-Fernàndez J, Gibson-Brown JJ, Grigoriev IV, Horton AC, de Jong PJ, Jurka J, Kapitonov VV, Kohara Y, Kuroki Y, Lindquist E, Lucas S, Osoegawa K, Pennacchio LA, Salamov AA, Satou Y, Sauka-Spengler T, Schmutz J, Shin-I T, Toyoda A, Bronner-Fraser M, Fujiyama A, Holland LZ, Holland PW, Satoh N, Rokhsar DS.* The amphioxus genome and the evolution of the chordate karyotype. *Nature.* 2008;**453**(7198):1064–71.
9. *Li ZB, Huang YS, Shangguan JB, Ning YF, Yi Y, Dai G.* Isolation and characterization of microsatellite loci in Branchiostoma belcheri Gray (Amphioxus). *Genet Mol Res.* 2015;**14**(3):10224–7.
10. *You L, Chi J, Huang S, Yu T, Huang G, Feng Y, Sang X, Gao X, Li T, Yue Z, Liu A, Chen S, Xu A.* LanceletDB: an integrated genome database for lancelet, comparing domain types and combination in orthologues among lancelet and other species. *Database (Oxford).* 2019;**2019**. pii: baz056.
11. *Yamada KD, Tomii K, Katoh K.* Application of the MAFFT sequence alignment program to large data-reexamination of the usefulness of chained guide trees. *Bioinformatics.* 2016;**32**(21):3246–3251.
12. *Nicholas KB, Nicholas HBJ.* GeneDoc: a tool for editing and annotating multiple sequence alignment. 1997
13. *Eddy SR.* SQUID-C function library for sequence analysis. 2008
14. *Rice P, Longden I, Bleasby A.* EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* 2000;**16**(6):276–7.
15. *Kramerov DA, Vassetzky NS.* Short retroposons in eukaryotic genomes. *Int Rev Cytol.* 2005;**247**:165–221.
16. *Kramerov DA, Vassetzky NS.* SINEs. *Wiley Interdiscip Rev RNA.* 2011;**2**(6):772–86.
17. *Albalat R, Martí-Solans J, Cañestro C.* DNA methylation in amphioxus: from ancestral functions to new roles in vertebrates. *Brief Funct Genomics.* 2012;**11**(2):142–55.

Надзвичайна різноманітність сімейств коротких ретропозонів (SINE) у ланцетника *Branchiostoma belcheri*

С. А. Косушкин, Н. С. Васецький

Короткі ретропозони (SINE)– важливі компоненти геному вищих еукаріот. Зазвичай в геномах є одне або кілька сімейств коротких ретропозонів. **Мета.** Пошук і аналіз SINE в геномі азіатського ланцетника (*Branchistoma belcheri*). **Методи.** Власні і стандартні методи комп'ютерного аналізу послідовностей. **Результати.** Було виявлено та проаналізовано 18 сімейств SINE B. belcheri. **Висновки.** Аналіз генома *B. belcheri* продемонстрував їх надзвичайна різноманітність. Обговорюється зв'язок великого числа сімейств коротких і довгих ретротранспозонов (LINE) ланцетників в зв'язку з особливостями геномів хордових.

Ключові слова: Короткі ретропозони, *Branchistoma belcheri*, ретротранспозони.

Чрезвычайное разнообразие семейств коротких ретропозонов (SINE) у ланцетника *Branchiostoma belcheri*

С. А. Косушкын, Н. С. Васецкий

Короткие ретропозоны (SINE) - важный компонент генома высших эукариот. Обычно в геномах есть одно или несколько семейств коротких ретропозонов. **Цель.** Поиск и анализ SINE в геноме азиатского ланцетника (*Branchistoma belcheri*). **Методы.** Собственные и стандартные методы компьютерного анализа последовательностей. **Результаты.** Было обнаружено и проанализировано 18 семейств SINE *B. belcheri*. **Выводы.** Анализ генома *B. belcheri* продемонстрировал их чрезвычайное разнообразие. Обсуждается связь большого числа семейств коротких и длинных ретротранспозонов (LINE) ланцетников в связи с особенностями геномов хордовых.

Ключевые слова: Короткие ретропозоны, *Branchistoma belcheri*, ретротранспозоны.

Received 02.09.2019