RESEARCH PAPERS

Identification of novel SINEs from Cyprinidae and their evolutionary significance*

HAN Yawei[1,2], HE Shunping[1**] and CHEN Yiyu[1]

(1. Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China. 2. Graduate School, Chinese Academy of Sciences, Beijing 100039, China)

Accepted on October 13, 2006

Abstract Short interspersed nuclear elements (SINEs) are widespread among eukaryotic genomes. They are repetitive DNA sequences that have been amplified by retrotransposition. In this study, a class of SINEs were isolated from the *Opsariichthys bidens* genome, and named *Opsar*. Sequence analysis confirmed that *Opsar* is a new class of typical SINEs derived from tRNA molecules. With the tRNA-derived region of *Opsar* and through BLASTN search, we further identified *Zb-SINE*s from the zebrafish genome, which includes two groups: *Zb-SINE*-A and *Zb-SINE*-B. The *Zb-SINE*-A group comprises subfamilies of -A1—-A5, and the *Zb-SINE*-B group is a dimer of the $tRNA^{Ala}$-derived region and shares a similar dimeric composition to Alu. *Zb-SINE*s are composed of three distinct regions: a 5′ end tRNA-derived region, a tRNA-unrelated region and a 3′ end AT-rich region. The flanking regions are AT rich. The average length of *Zb-SINE*s elements is about 340 bp. *Zb-SINE*s account for as much as 0.1% of the whole zebrafish genome. About 70% of the *Zb-SINE*s are on chromosomes 11, 18, and 19. These *Zb-SINE*s were characterized by PCR and dot hybridization. The distribution pattern of *Zb-SINE*s in genome strongly supports the master genes model. The tRNA-derived regions of *Opsar* and *Zb-SINE*s were compared with the $tRNA^{Ala}$ gene, and they showed 76% similarity, indicating that *Opsar* and *Zb-SINE*s originated from an inactive $tRNA^{Ala}$ sequence or a $tRNA^{Ala}$—like sequence. In view of the evolutionary status of zebrafish in the Cyprinidae, we deduced that *Zb-SINE*s were a very old class of interspersed sequences.

Keywords: short interspersed nuclear elements (SINEs), master gene model, *Opsar*, *Zb-SINE*s and Cyprinidae.

Short interspersed nuclear elements (SINEs) are a class of mobile repetitive molecules that have received much attention during the recent expansion of genomics. These molecules range from 70 to 500 bp in length[1]. There are usually 10^3—10^5 copies of a SINE family in a eukaryotic genome[2], such as human[3] and other mammals[4], amphibia[5], fish[6], fungi[7], and plants[8]. All SINEs characterized to date can be divided into three classes. The major class is SINEs derived from tRNA molecules[9], and the minor one includes SINEs derived from 7SL RNA and 5S rRNA; for example, the Alu, B1 and TuI and II families derived from 7SL RNA and present in primate, rodent and tree shrew genomes, respectively[10—12]; SINE3 in zebrafish genome derived from 5S rRNA[13]. Most SINEs utilize pol III promoters of type 2. However, SINE3 derived from 5S RNA utilizes type 1 of pol III promoter. Current evidence suggests that all SINEs reported to date are transcribed by pol III; then, the transcripts can be reversely transcribed and the DNA products will be inserted back into the host genome, simultaneously producing short direct repeats at the site of insertion[14]. Some SINEs have the ability to fuse with other sequences to produce composite transposable elements, for example, Alu occurs in monomeric, dimeric and sometimes tetrameric form[15,16]. SINEs are retrotransposons, an extremely successful class of mobile genetic elements as judged by their high copy numbers. However, most SINEs copies are transcriptionally silent. Nevertheless, the abundance and widespread presence of SINEs in eukaryotic genomes have a major impact on genome structure and function[17,18]. Two models, the master gene model and the multiple sources gene model, have emerged to explain the mechanism of SINE distribution in the genome[19,20]. Although the mechanism of SINE retrotransposition has not been completely understood

* Supported by the National Fund for Distinguished Young Scholars (Grant No. 30225008)

** To whom correspondence should be addressed: E-mail: clad@ihb.ac.cn

yet, many functions as well as some diseases have been proved to be associated with SINEs[21—23], and in view of the interest and activity in this area, more connections are likely to emerge in the near future.

Studies of cyprinid SINEs have almost exclusively concentrated on the zebrafish[13,24,25], with almost nothing reported about SINEs in other cyprinid species. Here, we report a class of SINEs derived from the $tRNA^{Ala}$ molecule from the *O. bidens* genome and the zebrafish genome, and named it *Opsar* and *Zb-SINE*s, respectively. PCR and dot hybridization confirmed that *Zb-SINE*s were specific components of the zebrafish genome, and the analysis of the *Zb-SINE*s sequences revealed two structural patterns, and *Zb-SINE*s amplification in the genome was putatively a master gene model.

1 Materials and methods

1.1 Materials

Specimens of *O. bidens* were colleted from River Yuanjiang (branch of Dongting Lake), Hunan, China and fixed in 95% alcohol. Specimens were identified in the Institute of Hydrobiology, CAS. Other Cyprinidae specimens, such as *Danio rerio*, *Mylopharyngodon piceus*, *Ctenopharyngodon idellus*, *Megalobrama amblycephala*, *Toxabramis swinhonis*, *Pseudobrama simony*, *Hypophthalmichthys molitrix*, *Aristichthys nobilis*, *Saurogobio dabryi*, *Gobiobotia filifer*, *Acheilognathus chankaensis*, *Acrossocheilus monticola*, *Garra orientalis*, *Cyprinus carpio*, *Schizothorax meridionnalis*, were from the Museum of Freshwater Fishes, Institute of Hydrobiology, CAS.

1.2 Extraction of genomic DNA, A-B PCR and inverse primer

Total genomic DNA was extracted from the muscles of specimens by a standard method[26]. A-B PCR was carried out as described by Borodulina and Kramerov[27]. Conditions for the A-B PCR were denaturation at 94℃ for 30 s, annealing at 34℃ for 30 s, and extension at 72℃ for 25 s, for 30 cycles. The PCR products were analyzed by electrophoresis in a 3% agarose gel. DNA bands were extracted and purified following manufacturer's protocol (Takara), cloned and sequenced. The 54 bp fragment from the A-B PCR was the basis of designing the reverse primers: BoxA: (5′-ACCGAGTCACCATCTCGCGTACG-3′) and BoxB: (5′-TATGAGGTCCTGGGTTCA ATCCC-3′).

1.3 Self-ligation of DNA fragments and SINEs amplification by inverse PCR

The genomic DNA was cleaved by *Hae* III. The 40 μL reaction mixture contained 22 μL 6 × buffer, 2.5 μg DNA, 2 units *Hae* III and 21 μL sterile double distilled water, and was incubated at 37℃ overnight. The *Hae* III was then inactivated by incubation at 70℃ for 1 h. The 10 μL sheared DNA fragments were added to 100 μL ligase reaction solution containing 4 μL T_4 ligase and 10 μL 10 × T_4 ligase buffer. The reaction was carried out at 16℃ overnight. The self-ligated DNA was purified, and dissolved in TE again. Inverse PCR was performed in a 50 μL reaction solution containing 0.5 unit *Taq* polymerase (Takara), 5 μL 2 × buffer, 4 μmol/L $MgCl_2$ and 2 μL of each primer (15 μmol) and 4 ng self-ligated DNA, and the reaction was repeatedly carried out at difference annealing temperature for farthest obtaining SINEs sequences except for denaturation at 94℃ for 30 s, extension at 72℃, and 30 cycles. PCR products were sequenced. Totally, 120 sequences were obtained. Five consensus sequences were constructed referencing Kapitonov and Jurk's method[13]. Then, the consensus sequences were analyzed according to SINEs characteristics.

1.4 Identification of *Zb-SINE*s in zebrafish genome

The tRNA-derived region of *Opsar* from *O. bidens* was searched by BLASTN in the HTGS (High Throughput Genomic Sequences) of zebrafish, and homologous copies from BLASTN in zebrafish genome were expanded ± 300 bp at their ends. The majority rule was applied to the multiple aligned expanded sequences from BLASTN, a preliminary multiple-sequence alignment was constructed by the program Clustal X[28]. The computer-generated alignment was subsequently carefully adjusted manually based on the Seaview program, at last, consensus sequences were obtained. Sequences were analyzed according to SINEs characteristics. All consensus sequences were recorded in GenBank as BX936356 (66111—421, 78667—965, 80102—425, 64704—5154, 78386—690, 62058—447). Their tRNA-derived regions were analyzed by the tRNAscan-SE program[29]. The *Zb-SINE*s copies in genome were identified as described by Kapitonov and Jurk[13].

1.5 The specificity of *Zb-SINE*s analyzed by PCR

The primers (5′-AAAAAAATRTTBVMCTT-TCAA-3′; 5′-TTMTTWTTCCGCCTTRTTASM G-3′) were designed according to *Zb-SINE*s sequences. PCR was performed in a 50 μL reaction volume containing 0.5 unit *Taq* polymerase (Takara), 5 μL 2× buffer, 4 μmol/L $MgCl_2$ and 2 μL of each primer (15 μmol) and 100 ng genomic DNA. Reaction conditions were denaturation at 94℃ for 30 s, annealing at 55℃ for 30 s, and extension at 72℃ for 60 s, for 30 cycles.

Dot hybridization was performed on the genomic DNA of representative species. *Zb-SINE*s from PCR amplification were labeled by DIG according to instruction manual (DIG high prime DNA labeling and detection starter Kit I; Roche). The 500 ng genomic DNA were used in dot hybridization.

1.6 Analysis of *Zb-SINE*s evolution

After alignment and analysis of *Zb-SINE*s, we analyzed the full length sequences by Bayesian 3.1 program[30] for explaining their phylogenetic relationships.

2 Results

2.1 Isolating *Opsar* from the *Opsariichthys bidens* genome

In this study, we firstly performed the inverse PCR for isolating the novel SINEs based on the method of A-B PCR[27], Ochman et al.'s method[31] and the typical characteristics of SINEs[32]. A new class of SINEs identified by inverse PCR from the *O. bidens* genome (experimental results not shown) was named *Opsar* for *Opsariichthys bidens*. The *Opsar* sequences are shown in Fig.1.

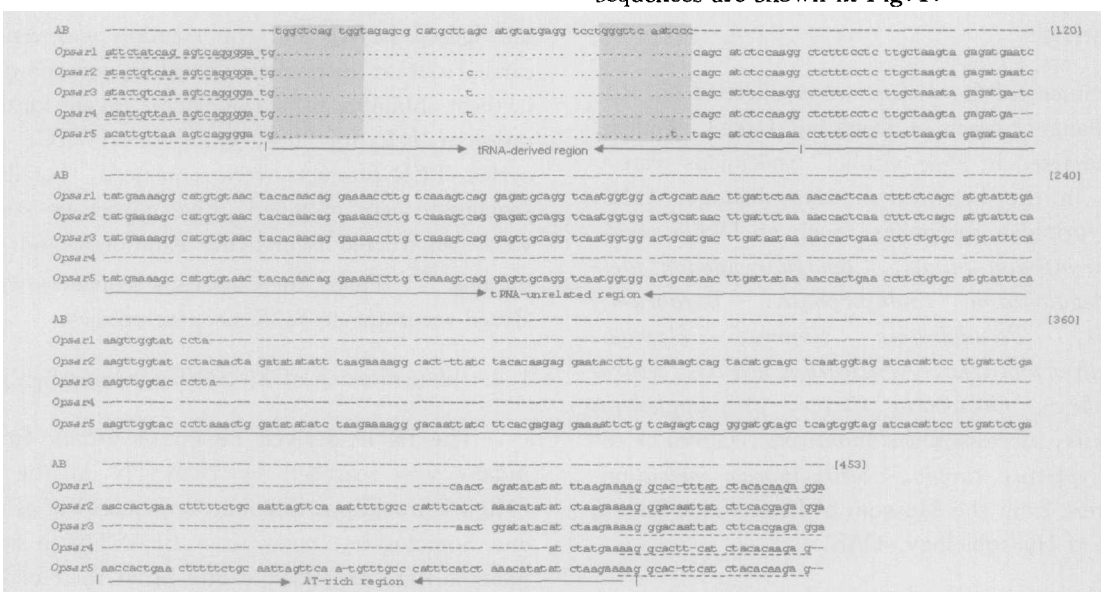

Fig. 1. The sequence of *Opsar*. The broken lines indicate the flanking regions and primers sequences. The pol III promoter boxes A and B are highlighted in grey. AB shows the results of A-B PCR. Dots indicate the same nucleotides, dashes show nucleotide deletions. The accession numbers of *Opsar*1—5 in GenBank are DQ075939, DQ075940, DQ075941, DQ075942 and DQ075943, respectively.

The 66 bp sequence is a tRNA-derived region at the 5′ end, about 65% similar to that of V-SINEs[25], which obviously displays A and B pol III promoter boxes. The length of the tRNA-unrelated region ranges from 38 to 248 bp in different *Opsar*. The tRNA-unrelated region of *Opsar* 4 is shorter than that of other *Opsar*. *Opsar* 2 and 5 have the longest tRNA-unrelated regions. The AT-rich region is located at the 3′ end. The AT-rich regions of *Opsar* 1, 3 and 4 are shorter than those of *Opsar* 2 and 5. The full lengths of *Opsar* are 149—452 bp, with *Opsar* 4 (149 bp) being the shortest. To prove that *Opsar* are SINEs from single self-ligated DNA fragments, we designed primers based on the flanking repeats, and PCR was preformed with the primers. The amplified sequences were cloned, sequenced and named *Ops*. After the *Ops* were aligned with the *Opsar* obtained by inverse PCR (data not shown), they proved identical. Our group also has successfully isolated some SINEs in other species by this method (These results will be reported in another paper), so that the reliability of method was confirmed. There-

fore, utilizing the inverse PCR method for isolating SINEs is much simpler than those previously reported[27,33—37].

2.2 Identification of *Zb-SINE*s in zebrafish genome

BLASTN was performed on the zebrafish genome using the sequences information from the tRNA-derived regions of *Opsar*. Novel repeats containing the $tRNA^{Ala}$-derived regions were identified. Using this observation as a starting point, we downloaded 130 similar zebrafish genome sequences. Applying the majority rule to a set of these sequences, we constructed six consensus sequences. When identifying these sequences, we found that they have some characteristics of random repeat. However, from their distribution in genome and sequence's characteristics, these sequences ought to be divided into SINEs family (see details in discussion). So we named it *Zb-SINE*s for zebrafish. *Zb-SINE*s comprise two sequence patterns in comparison with the typical SINEs and the tRNA-derived region. Fig. 2 shows a schematic representation.

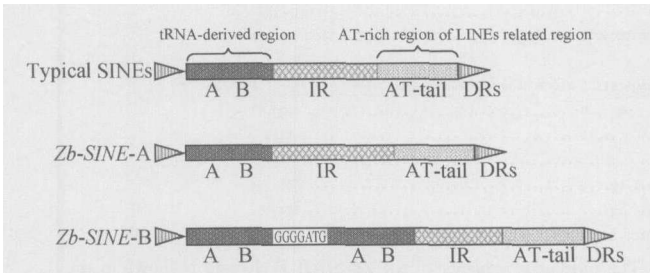

Fig. 2. Schematic structure of *Zb-SINE*s and the typical SINEs. A and B show the promoters of pol III; IR, internal regions or tRNA-unrelated regions; DR, the flanking repeats or target direct repeats. *Zb-SINE*-A and *Zb-SINE*-B are two typical models of *Zb-SINE*s.

Zb-SINE-A model comprises three regions: a 66 bp region similar to the $tRNA^{Ala}$ gene at the 5′ end, a 96—264 bp internal region, and an approximately 60 bp AT-rich region at the 3′ end. The *Zb-SINE*-B model is comprised of two tRNA similar regions which are linked by GGGGATG. Other regions in the *Zb-SINE*-B model are similar to those in *Zb-SINE*-A. Flanking repeats are located at both ends of the *Zb-SINE*s, and are AT-rich. All *Zb-SINE*s are shown in Fig. 3. To analyze *Zb-SINE*s distribution in the genome, *Opsar*1 was aligned with *Zb-SINE*s.

2.3 Localization of *Zb-SINE*s on chromosomes

The sequences from BLASTN clearly showed that *Zb-SINE*s were interspersed in the genome. By querying the zebrafish genome (http://www.ensembl.org/Danio_rerio/) using all the consensus sequences, we found that *Zb-SINE*s are distributed on chromosomes 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 13, 14, 16, 18, 19, 20, 21 and 23. Furthermore, they were mostly located at the chromosome ends and near the centromeres; especially, chromosomes 11, 18 and 19 contain all typical *Zb-SINE*s. Table 1 shows the chromosome locations and copy numbers of *Zb-SINE*s in the genome.

Table 1. Similar copies and chromosome location of *Zb-SINE*s in zebrafish genome

*Zb-SINE*s	Similar copy numbers in zebrafish genome	Chromosome location
Zb-SINE-A1	1064	1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 13, 14, 16, 18, 19, 20, 21, 23
Zb-SINE-A2	1065	1, 5, 7, 11, 18, 19, 20, 21
Zb-SINE-A3	1077	8, 11, 14, 18, 19, 21
Zb-SINE-A4	1982	1, 5, 7, 8, 11, 14, 18, 19, 20
Zb-SINE-A5	1028	5, 7, 8, 11, 14, 18, 19, 20
Zb-SINE-B	1881	1, 5, 7, 8, 11, 14, 18, 19, 20

Note: the above results were from the zebrafish genome in GenBank (http://www.ncbi.nlm.nih.gov/genome/guide/zebrafishindex.html) and EBI (http://www.ensembl.org/Danio_rerio/)

2.4 The 5′ ends of *Zb-SINE*s derived from the $tRNA^{Ala}$ molecule

The 5′ ends of *Zb-SINE*s were highly conserved, with 76% identity to the mouse $tRNA^{Ala}$ gene. Only two sites (C and T) showed mutations in the tRNA-derived region (Fig. 3). Usually, the tRNA gene exists in multiple copies in the genome. Four codons and three structures have been identified for alanine tRNA. Alignment of $tRNA^{Ala}$ with *Zb-SINE*s showed that the tRNA-derived regions of *Zb-SINE*s were similar to the $tRNA^{Ala}$ sequence around the anticodon TGC. We analyzed the *Zb-SINE*s tRNA-derived region using tRNAscan-SE Search Server (http://www.genetics.wustl.edu/eddy/tRNAscan-SE/). The results indicated that the anticondon was located at 36—38 bp in the tRNA-derived region. Their secondary structures also revealed similar characteristics. Most mutations have occurred in the amino acid acceptor stem, the TψC stem and the anticodon stem of the tRNA. Deletions have occurred in the D loop and the TψC loop, respectively, which are the pol III promoter sequences. Because the tRNA-derived regions of *Zb-SINE*s were obtained by

BLASTN of that of *Opsar*, these regions are the same. The secondary structures of the $tRNA^{Ala}$ gene and the consensus tRNA-derived region of *Opsar* and *Zb-SINE*s are shown in Fig. 4.

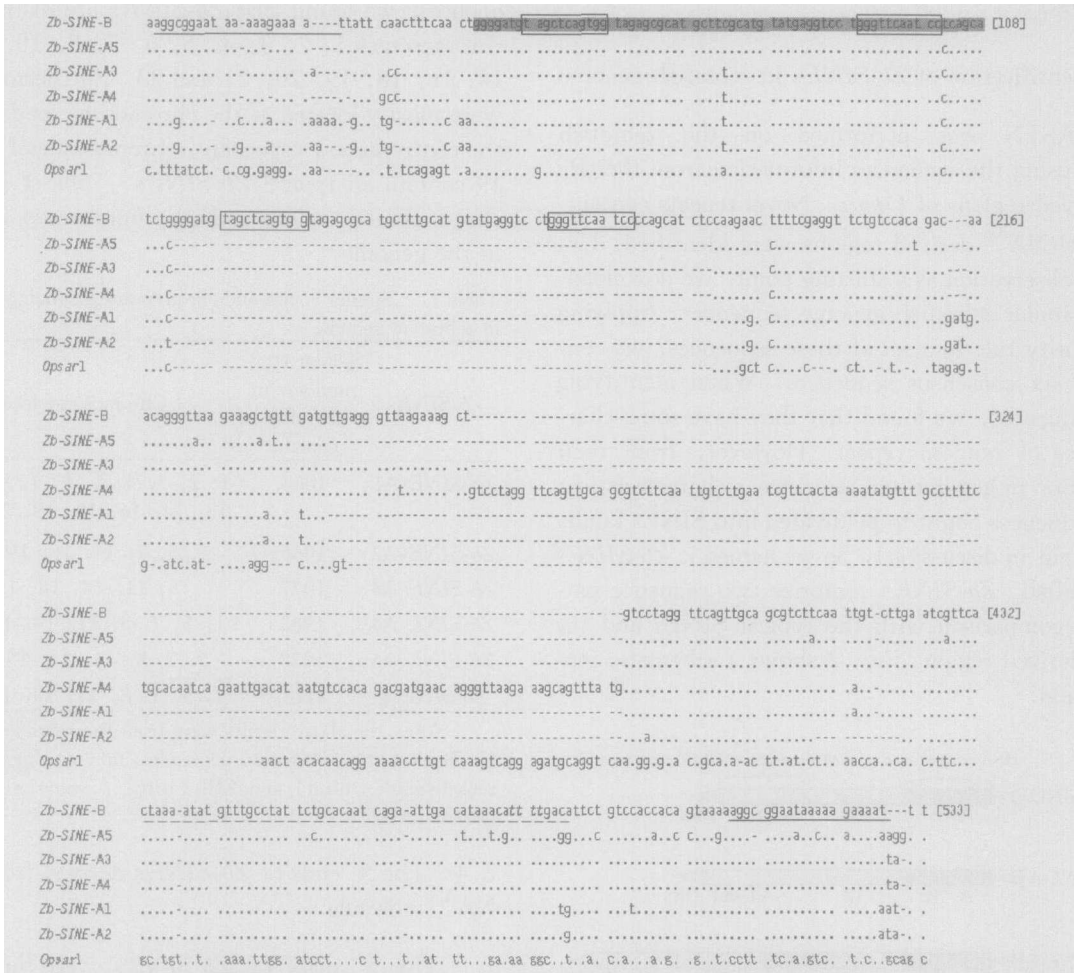

Fig. 3. The alignments of members of the *Zb-SINE*s and *Opsar*1. The consensus sequence of the *Zb-SINE*-B element is shown in the first line of the alignment. Dots indicate nucleotides identical to those in the consensus sequence. Dashes indicate insertions/deletions. The 5′ region derived from the $tRNA^{Ala}$ is highlighted in grey box. The sequences of pol III promoters are shown in blocks. The AT-rich region of *Zb-SINE*s is highlighted by a broken line. The flanking repeats of *Zb-SINE*s are underlined. The terminator is highlighted by italics. Their record numbers of *Zb-SINE*-A1 to -A5, and -B in GenBank sequences are BX936356 (66111—421, 78667—965, 80102—425, 64704—5154, 78386—690 and 62058—62447).

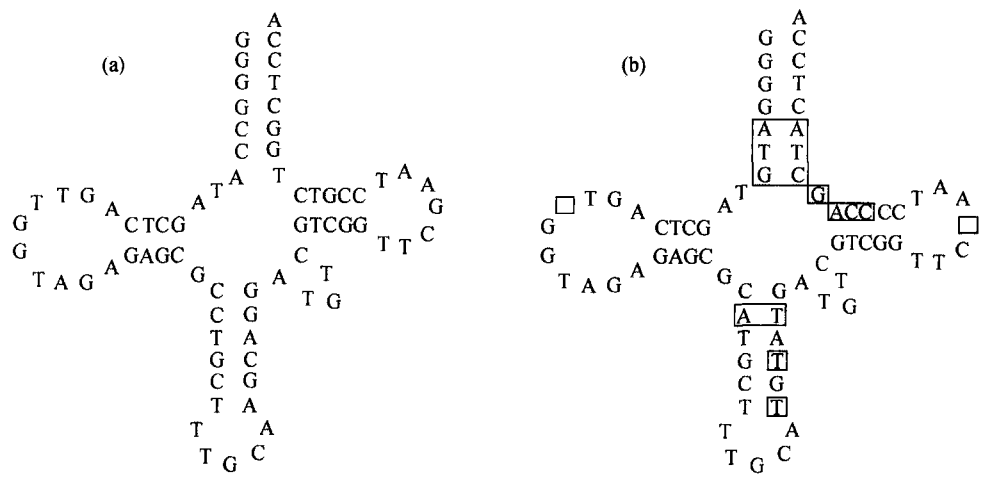

Fig. 4. The secondary structure of the $tRNA^{Ala}$ gene (a) and the tRNA-derived region of *Opsar* and *Zb-SINE*s (b). The blocks indicate the deletions. The letters in the blocks highlight the mutation sites.

2.5 The internal regions of *Zb-SINE*s

Internal regions or tRNA-unrelated regions of *Zb-SINE*s were highly variable. A deletion (110—178 bp) existed in the internal region of the *Zb-SINE*-A model. Although a few mutations and small deletions have also occurred in different *Zb-SINE*s subfamilies, the region of 180—238 bp was highly conserved in all *Zb-SINE*s. However, GAT was deleted in *Zb-SINE*-A3, -A4, -A5 and *Zb-SINE*-B. A long insertion of 260—385 bp was only present in *Zb-SINE*-A4. A short sequence (TTGAGGGTTAA-GAAAGCT) was inserted into *Zb-SINE*-A3 and *Zb-SINE*-B. *Zb-SINE*-A2 had a short deletion at 411—421 bp (CAATTGTCTT).

2.6 AT-rich regions and the flanking repeats of *Zb-SINE*s

The AT-rich region of *Zb-SINE*s was located at 432—498 bp in alignment (Fig. 3), which is conserved in all *Zb-SINE*s. The 3′ ends of *Zb-SINE*s are 75% AT, and contain a terminator (TTTT). By aligning the 3′ ends of *Zb-SINE*s and *ZfL2-1* that is an active retrotransposon in zebrafish genome and can encode reverse transcriptase[38], they have 70% similarity (Fig. 5).

```
ZfL2-1    actaaaacct gaa--cc-ac actgttccaa ttacta-tga ccattta--t gtgaagctgc tttgacacaa tctacattgt aaaa--gcgc tatacaaata aagct [105]
Zb-SINEs  .......ta. .tttg..t.t t...ca.-.. .c.ga.t... .-..aa.ca. t.--t.acat .c..t.-.- -.--..ca.. ....ag...g a...a...g. ..at.
```

Fig. 5. The alignment of the 3′ ends of *Zb-SINE*s and *ZfL2-1*. Dots indicate nucleotides identical to those in the ZfL2-1 sequence. Dashes indicate deletions.

From the above results, *Zb-SINE*s obviously show all the characteristics of SINEs. That is to say, *Zb-SINE*s are a new class of the typical SINE derived from the RNA^{Ala} gene in the zebrafish genome.

2.7 The specificity of *Zb-SINE*s confirmed in Cyprinidae

The primers from *Zb-SINE*s were used in PCR on the genomic DNA of all specimens. The results showed that *Zb-SINE*s only appeared in the zebrafish genome (Fig. 6). Furthermore, the multiple fragments in the PCR results were consistent with a diversity of *Zb-SINE*s length. To truly confirm the specificity of *Zb-SINE*s, the dot hybridization was performed on all representative species. The hybridization results also indicated that *Zb-SINE*s are of species-specific (Fig. 6).

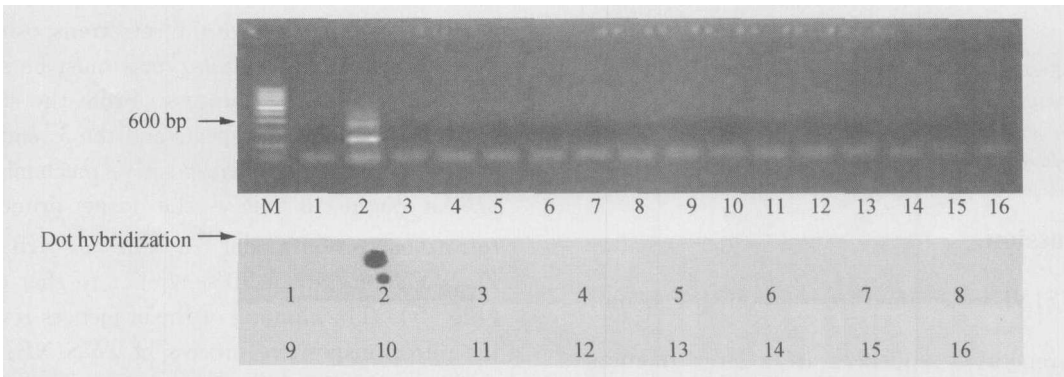

Fig. 6. The PCR results of *Zb-SINE*s and dot hybridization in Cyprinidae. M, DNA marker; 1, *Opsariichthys bidens*; 2, *Danio rerio*; 3, *Mylopharyngodon piceus*; 4, *Ctenopharyngodon idellus*; 5, *Megalobrama amblycephala*; 6, *Toxabramis swinhonis*; 7, *Pseudobrama simony*; 8, *Hypophthalmichthys molitrix*; 9, *Aristichthys nobilis*; 10, *Saurogobio dabryi*; 11, *Gobiobotia filifer*; 12, *Acheilognathus chankaensis*; 13, *Acrossocheilus monticola*; 14, *Garra orientalis*; 15, *Cyprinus carpio*; 16, *Schizothorax meridionnalis*.

2.8 *Zb-SINE*s evolution in the genome

Most investigations of SINEs amplification in genomes have focused on the relations of SINEs and LINEs, and copies of SINEs and SINEs activity[39—41]. In this study, we have utilized the sequence analysis method to prove the amplifying model of *Zb-SINE*s in the zebrafish genome. The results are shown in Fig. 7. *Opsar*1 from *O. bidens* genome was assigned as out-group. From the results, we may know that *Zb-SINE*-A1 and -A2 are derived from a common ancient sequence; the insertion of GAT only happens in the internal region of *Zb-SINE*-A1 and -A2; and the 10 bp deletion (CAATTGTCTT) distinguishes the *Zb-SINE*-A1 from *Zb-SINE*-A2. *Zb-SINE*-A4 is in the sister group of *Zb-SINE*-A1 and -A2 ancient sequences. *Zb-SINE*-A3 and-A4 sequences have a 98% homology, except for an inser-

tion of 126 bp (260—385 bp) in the internal region. However, sequence analyses indicated that *Zb-SINE*-A3 was older than *Zb-SINE*-A4 in evolution (Fig. 7). *Zb-SINE*-B and *Zb-SINE*-A3 have a close relation in evolution, and derived a common ancient sequence. *Zb-SINE*-A5 has a short insertion (TTGAGGTTAAGAAAGCT) relative to *Zb-SINE*-B in the internal region. Though *Zb-SINE*-B is a dimmer of the tRNA-derived sequence, *Zb-SINE*-A5 is older than *Zb-SINE*-B in evolution. The bootstrap rates (0.62—0.93) also exhibited that the total evolutionary relationship is supported. From the total evolutionary model, the ancient sequence of *Zb-SINE*-A1 and -A2 can propagate the offspring, whereas, the *Zb-SINE*-A4 does not have the progenitive ability. The similar model goes down to *Zb-SINE*-A5. This evolutionary relationship clearly supports the master genes model. In other words, the ancestral sequences of *Zb-SINE*s only produce a prototype sequence that generated offspring in each process of retrotransposition.

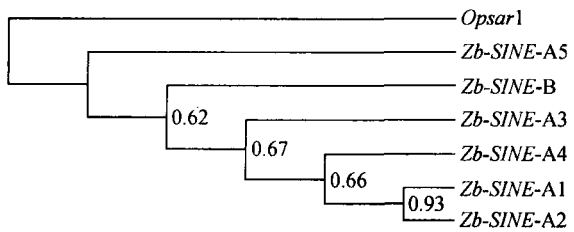

Fig. 7. Evolutionary relationship of the *Zb-SINE*s analyzed using Bayesain method. The number of substitution types (nst): 6GTR; the number of generations: 10^6. *Zb-SINE*-A1, -A2, -A3, -A4, and -A5 belong to the *Zb-SINE*-A pattern. *Zb-SINE*-B is the *Zb-SINE*-B pattern.

3 Discussion

3.1 *Zb-SINE*s identified from zebrafish genome

The repetitive sequences are a large fraction of genome, and SINEs are belonging to one of the repetitive families. When studying SINEs, it is an essential question that how to identify novel SINEs from genome. In 2004, Okada et al. introduced the method of isolating and identifying novel SINEs[37]. In fact, identifying SINEs is based on three principles: (1) all SINEs known to date have a region transcribed by pol III; (2) SINEs contain the AT-rich tail; (3) usually, SINEs have a target site duplication. Of course, some SINEs are exceptive, because Weiner et al. reported that insertion of a nonviral retroposon usually generated a target site duplication of 7—21 bp, but shorter and longer direct repeats or none at all[42]. In this study, *Zb-SINE*s also exhibited some characteristics of the random repeat in some chromosome, for instance, some region (BX936356 contig) from chromosome 21. However, from the chromosome locations and *Zb-SINE*s characteristics, *Zb-SINE*s should belong to the SINEs family, for they are distributed on all the chromosomes and contain the tRNA-derived region. Although the flanking regions of *Zb-SINE*s seemed longer than common SINEs, the clustering distribution of SINEs in the genome can explain the phenomenon[28,43]. In addition, this distribution may result from the unequal homologous recombination or exchange[44]. Therefore, we concluded that *Zb-SINE*s are a typical SINEs family in the zebrafish genome.

3.2 The retrotranspositive mechanism of *Zb-SINE*s

Most of the enzymes directly involved in SINEs retrotransposition are still unknown. However, some convincing data indicated that SINEs amplification in host genome was catalyzed *in trans* by the reverse transcriptiase encoded by LINEs[38]. In this study, it is clear that the *Zb-SINE*s are a group of typical SINEs derived from $tRNA^{Ala}$ in the zebrafish genome. Although it is unknown whether *Zb-SINE*s retrotransposition is related to that of non-LTR or whether *Zb-SINE*s have the retrotranspositive activity, it is clear that some enzymes must be involved in the retrotranspositive process. From the structure of the AT-rich flanking repeats and the 3′ end, we conjecture that the retrotranspositive mechanism of *Zb-SINE*s ought to follow the target-primed reverse retrotransposition model[16]. The AT-rich regions of *Zb-SINE*s are about 70% similar to that of *ZfL2-1* (Fig. 5). The aligning of the sequences revealed that the retrotranspositive process of *Zb-SINE*s might be mediated by *ZfL2-1*, because the previous experiment had confirmed that the *ZfL2-1* was a kind of LINEs, and could mediate SINEs retroposition in zebrafish genome[38]. Furthermore, some research groups had confirmed that the 3′ tails of tRNA-derived SINEs were derived from that of long interspersed repetitive elements (LINEs)[38,39].

The 5′ end of *Zb-SINE*-B is a dimer of the similar $tRNA^{Ala}$ gene. Dimeric structures such as *Zb-SINE*-B have previously been reported[20,37], and there is indication that they may be generated in a variety of ways[45]. A dimeric retrotransposon structure

may result from the duplication of a tRNA gene. Alternatively, an original gene cluster of two functional tRNA genes may be responsible for the dimeric structure[20]. This possibility had also been discussed for the two arginine tRNA-derived monomers of the *Twin* SINE family from the vector mosquito *Culex pipiens*[46]. Analysis of the evolutionary relationship among the *Zb-SINE*s suggested that when *Zb-SINE*-A3 or -A4 was amplified by retrotransposition, the cDNA sequences of *Zb-SINE*-A3/-A4 or the ancestral sequences of *Zb-SINE*s -A3 and -A4 were inserted into the 3′ end of the tRNA-derived region, thus bringing up the *Zb-SINE*-B structure, because the same short GGGGATG also appears at the 5′ end of the second tRNA-derived region. The internal region of *Zb-SINE*-A4 has a long inserting sequence similar to the 3′ end sequences of other *Zb-SINE*s. This may have been the result of duplication of *Zb-SINE*-A3 ancient sequence mRNA. When the mRNA of the ancient sequence was duplicating, its 3′ end was duplicated twice because of the action of certain factors. At present, although we cannot prove by experiments whether this ancient mRNA hypothesis is correct, it seems the most plausible explanation for the production of *Zb-SINE*-A4. Other short deletions or insertions in the *Zb-SINE*s may have been produced during duplication or interspersed among the host genes.

3.3 The chromosome distribution of *Zb-SINE*s

The chromosome distribution of *Zb-SINE*s shows that they are mostly on chromosomes 11, 18, and 19. The zebrafish genome contains 25 chromosomes comprising 17418 genes recorded to our study by EBI (http: //www.ensembl.org/Danio_rerio/). Only 11% of the total genes are contained by chromosomes 11, 18, and 19. However, about 70% of the copies of *Zb-SINE*s are distributed among these 3 chromosomes; in particular, chromosomes 11 and 19 contain all *Zb-SINE*s types. Therefore, we believe that these three chromosomes are rich in short interspersed elements.

3.4 The species-specificity of *Zb-SINE*s and the relation of *Opsar* and *Zb-SINE*s

Because *Zb-SINE*s were identified from the tRNA-derived region of *Opsar*, *Opsar* and *Zb-SINE*s have the same tRNA-derived region at 5′-ends. However, from the PCR results, we know that *Zb-SINE*s have only appeared in the zebrafish genome. The dot hybridization results reached an identical conclusion. These indicate that *Zb-SINE*s have a species-specific nature. The multiple fragments revealed by electrophoresis were also consistent with a range of *Zb-SINE*s lengths. Because *Danio rerio* and *Opsarichthys biden* were classified as Danioninae, a very old taxon[47], they were evolutionarily old. Since *Zb-SINE*s were obtained on the basis of *Opsar*, when we analyzed the evolution of *Zb-SINE*s, there were theoretical grounds for sustaining *Opsar* 1 as outgroup. Other species do not give positive results in PCR and dot hybridization test. Therefore, we inferred that the *Zb-SINE*s were interspersed in the genome before the early Tertiary.

Another question we tried to address was whether *Opsar* and *Zb-SINE*s were evolutionarily related. Comparison of the $tRNA^{Ala}$ gene with the tRNA-derived regions of *Opsar* and *Zb-SINE*s led us to conjecture that *Opsar* and *Zb-SINE*s may originate from an inactive $tRNA^{Ala}$ or a $tRNA^{Ala}$-like sequence recombined with retrotranspositive elements. Differences between the $tRNA^{Ala}$ molecule and the tRNA-derived regions of *Opsar* and *Zb-SINE*s are apparent in their secondary structures (Fig. 4). There are also deletions in the loop of the D and the TψC loop, which are the pol III promoter sequences. These deletions may have resulted in tRNA gene inactivity and *Opsar* and *Zb-SINE*s came into being.

3.5 The evolutionary model of *Zb-SINE*s in genome

Two major models have been proposed to explain the amplification of SINEs in the genome: the master gene model[40,48] and the multiple-source genes model[19,20]. In the master gene model, only a single or a few "master gene" loci give rise to all offspring copies. The multiple-source gene model includes offspring that can propagate in the same manner as the parent copies, thus serving as "multiple-source genes" over evolutionary time. The target-primed reverse transcription (TPRT) may explain SINEs amplifying mechanism; this was the first hypothesis advanced to explain the relationship between SINEs and LINEs[16]. To elucidate the evolutionary model of *Zb-SINE*s, their entire length sequences were analyzed using a Bayesian (3.0) program[31]. The evolutionary relationships of *Zb-SINE*s clearly indicate that *Zb-SINE*s-A5 and -B are older than the others, however, *Zb-SINE*-A5 does not show the progenitive ability, and only the sister group of *Zb-SINE*-A5 can

propagate the offspring. The same model can be extrapolated to *Zb-SINE*-A1, as shown in Fig. 7. Only the offspring of *Zb-SINE*s seem to have been reproduced from the ancestral sequence. The master genes model is clearly corroborated by the whole evolutionary tree. Although we have no way to determine whether this deduction is correct, it is worthful that the computational result was obtained from statistical analysis by the Bayesian program. From the structure of the flanking regions, *Zb-SINE*-A5 has more variation than others. Two issues arise. Firstly, when *Zb-SINE*s were inserted into the host genome, two flanking repeat patterns were generated. This characteristic indicates their evolutionary differences. Secondly, their structural divergence may have resulted from different enzymes mediating retrotransposition in the genome. Their evolutionary relationship indicated that *Zb-SINE*-A5 was propagated from a sister group of *Zb-SINE*-B.

3.6 *Zb-SINE*s may be utilized to study Cyprinidae evolution

Cypinidae, the largest fish family, comprises approximately 210 recognized genera and 2010 species that are distributed widely in Eurasia, East Indian Islands, Africa and North America[49]. From the evolutionary history, the Danioninae are the oldest taxon. According to the fossil record, geological time and geographic distribution, Danioninae had already existed in the early Tertiary[47]. Part of the Danioninae genus was also rearranged into the primitive Tertiary group by He et al[50]. These data indicate that the Danioninae taxon is a very old and has an important status in evolution. *O. bidens* only exists in Eastern Asia. Zebrafish is a species of Danioninae. SINEs can provide the excellent evolutionary information for reconstructing phylogenesis and have already been used as molecular markers in systematics[17,28]. Therefore, *Opsar* and *Zb-SINE*s might prove to be molecular markers for the phylogeny of Cyprinidae. At the same time, this work has also provided a rapid and simple way of isolating SINEs from a genome.

In summary, we have developed an experimental method for isolating SINEs from genome, and obtained the full length sequence of *Opsar* by inverse PCR on the genomic DNA. A group of SINEs in the zebrafish genome was identified by bioinformatics methods, and named *Zb-SINE*s. First, we observed that *Zb-SINE*s are a class of the typical SINEs derived from $tRNA^{Ala}$. Second, *Zb-SINE*s comprise two subfamilies (A and B patterns) differing in their tRNA-derived region structures. Third, *Zb-SINE*s exist in most zebrafish chromosomes and constitute 0.1% of the zebrafish genome. Fourth, we have shown that the amplification pattern of *Zb-SINE*s in the genome supports the master genes model. Finally, the specific nature of *Zb-SINE*s in Cyprinidae has been confirmed.

Acknowledgements We are very grateful to Prof. Manyuan Long (University of Chicago, USA) for revision of the manuscript, and to Tong Chaobo and Chen Liping for assistances in the experiment.

References

1 Shedlock A. M., Takahashi K. and Okada N. SINEs of speciation: tracking lineages with retroposons. Trends Ecol. Evol., 2004, 19: 545—553.
2 Singer M. F. SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. Cell, 1982, 28: 433—434.
3 Ray D. A., Walker J.A., Hall A. et al. Inference of human geographic origins using Alu insertion polymorphisms. Forensic Sci. Int., 2005, 153: 117—124.
4 Jinchuan X., Hui W., Kyudong H. et al. A mobile element based phylogeny of old world monkeys. Mol. Phylogenet. Evol., 2005, 37: 872—880.
5 Shimoda N., Chevrette M., Ekker M. et al. Mermaid: a family of short interspersed repetitive elements widespread in vertebrates. Biochem. Biophys. Res. Comm., 1996, 220: 226—232.
6 Takahashi K. and Okada N. Mosaic structure and retropositional dynamics during evolution of subfamilies of short interspersed elements in African Cichlids. Mol. Biol. Evol., 2002, 19: 1303—1312.
7 Kachroo P., Leong S. A. and Ghattoo B. B. Mg-SINE: a short interspersed nuclear element from the rice blast fungus, Magnaporthe grisea. Proc. Natl. Acad. Sci. USA, 1995, 92: 11125—11129.
8 Thierry P., Cécile B. A., Laurence L. et al. Synthesis and processing of tRNA-related SINE transcripts in *Arabidopsis thaliana*. Nucleic. Acids. Res., 2004, 32: 3957—3966.
9 Ogiwara I., Miya M., Ohshima K. et al. Retropositional parasitism of SINEs on LINEs: identification of SINEs and LINEs in elasmobranches. Mol. Biol. Evol., 1999, 16: 1238—1250.
10 Kramerov D. A., Lekakh I. V., Samarina O. P. et al. The sequences homologous to major interspersed repeats B1 and B2 of mouse genome are present in mRNA and small cytoplasmic poly $(A)^+$RNA. Nucleic. Acids. Res., 1982, 10: 7477—7491.
11 Ullu E. and Tschudi C. Alu sequences are processed 7SL RNA genes. Nature, 1984, 312: 171—172.
12 Nishihara H., Terai Y. and Okada N. Characterization of novel Alu- and tRNA-related SINEs from the Tree Shrew and evolutionary implications of their origins. Mol. Biol. Evol., 2002, 19: 1964—1972.
13 Kapitonov V. V. and Jurka J. A novel class of SINE elements derived from 5S RNA. Mol. Biol. Evol., 2003, 20: 694—702.
14 Smit A. F. The origin of interspersed repeats in the human genome. Curr. Opin. Genet. Dev., 1996, 6: 743—748.

15 Lee M. G., Loomis C. and Cowan N. J. Sequence of an expressed human beta-tubulin gene containing ten Alu family members. Nucleic. Acids. Res., 1984, 12: 5823—5836.
16 Luan D. D., Korman M. H., Jakubczak J. L. et al. 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.
17 Khitrinskaya I. Y., Stepanov V. A. and Puzyrev V. P. Alu repeats in the human genome. *Mol Biol* 2003, 37: 325—333.
18 Jurka J. Evolutionary impact of human Alu repetitive elements. Curr. Opin. Genet. Dev., 2004, 14: 603—608.
19 Shedlock A. M. and Okada N. SINE insertions: powerful tools for molecular systematics. BioEssays, 2000, 22: 148—160.
20 Piskurek O., Nikaido M., Boeadi M. et al. Unique mammalian tRNA-derived repetitive elements in dermopterans: the t-SINE family and its retrotransposition through multiple sources. Mol. Biol. Evol., 2003, 20: 1659—1668.
21 Deininger P. L. and Batzer M. A. Alu repeats and human disease. Mol. Genet. Metab., 1999, 67: 183—193.
22 Callinan P. A., Wang J., Herke S. W. et al. Alu retrotransposition-mediated deletion. J. Mol. Biol., 2005, 348: 791—800.
23 Price A. L., Eskin E. and Pevzner P. A. Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. Genome Res., 2004, 14: 2245—2252.
24 Zsuasanna I., Aolatán I., David G-E. et al. DANA elements: A family of composite, tRNA-derived short interspersed DNA elements associated with mutational activities in zebrafish (Danio rerio). Proc. Natl. Acid. Sci. USA, 1996, 93: 1077—1081.
25 Ogiwara I., Miya M., Ohshima K. et al. V-SINEs: a new superfamily of vertebrate SINEs that are widespread in ertebrate genomesand retain a strongly conserved segment within each repetitive unit. Genome Res., 2002, 12: 316—324.
26 Sambrook J., Fritsch E. F. and Maniatis T. Molecular Cloning: A Laboratory Manual, 2nd ed. New York: Cold Spring Harbor Laboratory Press, 1989.
27 Borodulina O. R. and Kramerov D. A. Wide distribution of short interspersed elements among eukaryotic genomes. FEBS Letters., 1999, 457: 409—413.
28 Thompson J. D., Gibson T. J., Plewniak F. 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.
29 Lowe T. M. and Eddy S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic. Acids. Res., 1997, 25: 955—964.
30 Ronquist F. and Huelsenbeck J. P. MrBayes 3: Bayesian phylogenetic inferenceunder mixed models. Bioinformatics, 2003, 19: 1572—1574.
31 Ochman H., Gerber A. S. and Hartl D. L. Genetic applications of an inverse polymerase chain reaction. Genetics, 1988, 120: 621—623.
32 Hillis D. M. SINEs of perfect character. Proc. Natl. Acad. Sci. USA, 1999, 96: 9979—9981.
33 Borodulina O. R. and Kramerov D. A. PCR-based approach to SINE isolation: simple and complex SINEs. Gene, 2005, 349: 197—205.
34 Kramerov D. A., Grigoryan A. A., Ryskov A. P. et al. Long double-stranded sequences (dsRNA-B) of nuclear pre-mRNA consist of a few highlyabundant classes of sequences: evidence from DNA cloning experiments. Nucleic. Acids. Res., 1979, 6: 697—713.
35 Krayev A. S., Markusheva T. V., Kramerov D. A. et al. Ubiquitous transposon-like repeats B1 and B2 of the mouse genome: B2 sequencing. Nucleic. Acids. Res., 1982, 10: 7461—7475.
36 Bennett K. L., Hill R. E., Pietras J. M. et al. Most highly repeated dispersed DNA families in the mouse genome. Mol. Cell. Biol., 1984, 4: 1561—1571.
37 Okada N., Shedlock A. M. and Nikaido M. Retroposon mapping in molecular systematics. Methods Mol. Biol., 2004, 260: 189—226.
38 Sugano T., Kajikawa M. and Okada N. Isolation and characterization of retrotransposition-competent LINEs from zebrafish. Gene, 2005, 365: 74—82.
39 Okada N., Hamada M., Ogiwar I. et al. SINEs and LINEs share common 3′ sequences: a review. Gene, 1997, 205: 229—243.
40 Deininger P. L. and Batzer M. A. SINE master genes and population biology. In: The Impact of Short Interspersed Elements (SINEs) on the Host Genome, edited by Maraia R. Austin, Texas: R.G. Landes Company, 1995, 43—60.
41 *Ohshima K., Hamada M., Terai Y. et al.* The 3′ ends of tRNA-derived short interspersed elements are derived from the 3′ ends of long interspersed repetitive elements. Mol. Cell. Biol., 1996, 16: 3756—3764.
42 Wenier A. M., Deininger P. L. and Efstratiadis A. Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu. Rev. Biochem., 1986, 55: 631—661.
43 Jurka J., Kohany O., Pavlicek A. et al. Clustering duplication and chromosomal distribution of mouse SINE retrotransposons. Cytogenet. Genome Res., 2005, 110: 117—123.
44 Kramerov D. A. and Vassetzky N. S. Short retroposons in eukaryotic genomes. Int. Rev. Cytol., 2005, 247: 165—221.
45 Rogers J. H. The origin and evolution of retroposons. Int. Rev. Cytol., 1985, 93: 187—279.
46 Feschotte C., Fourrier N., Desmons I. et al. Birth of a retroposon: the Twin SINE family from the vector mosquito culex pipiens may have originated from a dimeric tRNA precursor. Mol. Biol. Evol., 2001, 18: 74—84.
47 Chen Y. Y. General review to Cyprinidae, in Fauna Sinica, Osteichths: Cypriniformes, Part II (In Chinese). Edited by Chen Y. Beijing: Science Press. 1998, pp. 1—18.
48 Kim J., Martignetti J. A., Shen M. R. et al. Rodent BC1 RNA gene as a master gene for ID element amplicifcation. Proc. Natl. Acad. Sci. USA, 1994, 91: 3607—3611.
49 Nelson J. S. Fishes of the World, 3rd ed. New York: John Wiley and Sons, 1994.
50 He S. P., Liu H. Z., Chen Y. Y. et al. Molecular phylogenetic relationships of Eastern Asian Cyprinindae (Pisces: Cypriniformes) inferred cytochrome b sequences. Science in China, 2004, 47: 130—138.