**THREE SUBFAMILIES OF SHORT INTERSPERSED ELEMENTS OF THE DOG GENOME: POSSIBLE ORIGIN AND FUNCTIONS.**

KOLESNIKOV N.N.+, ROGOZIN I.B., LAVRENTIEVA M.V., ELISAPHENKO E.A.

Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, 10 Lavrentiev Ave., Novosibisk, 630090, Russia;
e-mail: kolesnik@cgi.nsk.su;

+Corresponding author

Keywords: dog genome, interspersed repetitive elements, multiple alignment, consensus

56 new repetitive elements, were found in the dog genome DNA sequences from databases. They have a typical general structure of mammalian SINEs, but also they possess some distinguished features, as a purine rich body and a microsatellite region. A multiple alignment showed that there are at least three subfamilies of SINE in the dog genome, differing point mutations, deletions and insertions. A parental gene, tRNA Lys, was established. Repetitive sequences localize in various gene regions (5' 3'-UTR, introns) and in some cases possibly participate in gene expression. In two cases parts of elements were found in coding gene region. Time of divergence of these three subfamilies and SINEs of mink and seal genomes was estimated.

Different classes of repetitive DNA elements varying in the size, organization and copy number have been revealed in mammalian genomes. Short and long interspersed repetitive elements (SINEs and LINEs) are the most abundant. The absences of the obvious specific function of these repeats led to term them as a selfish DNA. The recent discoveries showed that they are not useless part of genome but may interact with the whole genome and play an important role in its evolution and function (for review see [1]).

To study these questions, the distribution and evolution of SINEs in the dog genome were examined. First short interspersed elements in the dog genome were found by Minnick et al. [2], similar repeats (called B-2 like) were described before in the mink genome [3] and later in the seal genome [4]. They were called Can SINE, as all three species belong to the superfamily Canoidea. It was suggested that they derived from tRNA alanine and their distribution restricted to canoid carnivores [4].

56 repetitive elements (reduced and full sized) were found in NCBI NIH (non-redundant database). Computer analysis was made by using BLAST program [5] and FASTA [6]. A multiple alignment of these sequences was made by using CLUSTAL program [7], consensus sequence was generated basing on the multiple alignments.

The general structure of the Canoidea SINEs family based on 58 repeats of dog*(Canis familiaris),*8 repeats of mink*(Mustela vison)*and 7 repeats of seal*(Phoca vitulina)*genomes are shown in Fig. 1. 12 homologous sequences were found in chicken genome.

 1 100 150-170 210 250

 tRNA like head Py-rich body MS A-rich tail

 R \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_-------------\_\_\_\_\_\_\_\_\_------------ R

 CT - 75% AA(TAAA)n...

 A box B (CT)n/ ..TC(T)n..

 /(TG)n ...(A)n

Figure 1. General structure of the Canoidea SINE. R - host target DNA repeats (7-20 bp in length); A,B - internal promoter regions for RNA-polymerase III; MS - microsatellite region; AA(TAAA)n -overlapping polyadenylation signal; TC(T)n - pol III termination signal. The size of the main parts of the consensus sequence is shown in bp above.

A total length of elements varies in size from 135 to 253 bp. The variation is mainly due to Py-rich body, MS - region and A-rich tail. On the basis of multiple alignment of 58 repeats three consensus sequences were revealed in the dog genome. Probably they represent three subfamilies (Cf1-3) which are distinguished by point mutations, insertions and deletions in diagnostic positions (Fig.2). The investigation of sequence divergence (discussed below) suggests that the Cf1 is more ancient, it has a very important insertion (position 36-43), what is absent in Cf2,Cf3 and in the SINEs of seal and mink. This sequence contains anticodon for tRNA Lys, which was probably lost by others subfamilies. The comparison of the dog consensus with the mammalian tRNA sequences showed the greatest similarity (74%) to a mouse (DK8101) and human (DK9991) tRNA lysine, suggesting that the last is a parent gene. All subfamilies contain 19 bp insertion after anticodon what is absent in mammalian tRNA Lys genes, but mollusc genes encoding lysine tRNA contain 19bp intron (Fig. 2). It was very suprising to find similar repeated elements in the chicken genome, tRNA-related and tRNA-unrelated parts of which are highly homologous with Cf1 (Fig 2). The origin of the tRNA unrelated parts is still an open question.

It was interesting to analyze localization of members of SINE family with respect to transcribed regions in the dog genome. It appears that 50% elements are located in introns, 20% in 5'UTRs and 30% in 3’UTRs. In this connection some structural features of the Canoidea SINEs may imply their possible specific function. It is microsatellite dinucleotide repeats (Fig.1). It's known that (CT)n - oligonucleotides are capable of forming unusual DNA conformation and in a such way play a role in transcriptional regulation. There is also a signal of polyadenylation, what may be functional in the case of insertion in plus orientation in 3'UTR of a gene. Moreover, we found short fragments of Cf1 elements (24 and 52 bp) included in the coding parts of interferon alpha gene and vascular anastomoticupregulated protein m-RNA with homology 75 and 71% respectively. Summarizing all data what are known at present about SINEs perhaps, we have to rename it as altruistic DNA.

 1 10 20 30 40 50

 \* A box \* \*\*

 a b b c

 ------> ---> <--- ----> lys

 Cf1 GGGGGCGCCTGGG TGGCTCAGTTGG TTAAGCATCTGACTCTTCGCTTTTG

 Cf2 GGGGRTRCCTGGG TGGCTCAGT-GG TTGAGCATCT--------GCCTTTG

 Cf3 GGGGATCCCTGGG TGGCGCAG-CGG TTTGGCGCCT--------GCCTTTG

 Mv --GGGCGCCTGGG TGGCTCAGTGGG TTAAGCCGCT--------GCCTTCG

 Pv GGGGTGGCCTGG- TGGCTCAGTCGG TTAAGNNNCT--------GCCTTCG

 Gg GGGGGCACCTGGG TGGCTCAGTCGG TTAAGCGTCCGACTCTT-GGTTTCA

tRNA LYS:

 DK8101 GCCCGGC TAGCTCAGTCGG TAGAGCATGAGACTCTTA-------

 D50536 GCCTCCA TAGCTCAGTCGG TAGAGCATCAGACTCTTAAGTATAC

 52 \* B box \* 100

 c d d a

 <---- ------> <-----<------

 Cf1 GCTCAGGTCATGATCTCAGGGTTG-TR AGATCRAGCCC YACATCGGGCT

 Cf2 GCTCAGGGCATGATCCYGGGGTCC-TG GGATCGAGTCC CRCATCRGGCT

 Cf3 GCCCAGGGCRYGATCCTGGAGACC-CG GGATCGARTCC CACGTCGGGCT

 Mv GCTCAGGTCATGATCTCAGGGTCC-TG GGATCGAGTCC CGCATCGGGCT

 Pv GCTCAGGTCATGATCTCAGGRTCC-TG GGATCGAGTCC CGCATCGGGCT

 Gg GCTCAGGTCATGATCTCANGGTTTGTG AGATCGAGCCC YRCATCGGGCT

 DK8101 ------------ATCTCAGGGTCG-TG GGTTCGAGCCC CACGTTGGGCG

 D50536 GTCGCATAAGAAATCTGAGGGTCT-GG GGTTCGAGTCC CCATGTGGGCT

 104 127 155

 Cf1 CCCYRCT--CAGYGGGGAGTCTGCTTGAGATTCTCTCYCYYTCTCCCTCTGCCT

 Cf2 CCCTG-------CAGGGAGCCTGCTT----------------CTCCCTCTGCCT

 Cf3 CCCTG-TG----CATGGAGCCTGCTT----------------CTCCCTCTGCCT

 Mv CTCTGCT--CRGCAGGGAGCCTGCTTCTCTTCCCTCCTCTCTCTCTCTCTGCCT

 Pv CCCTGCT--CAGCAGGGAGCCTGCTT----------CTCCCTCTCCCTCTGCCT

 GgN14 CTGYGCTGACAGTGTGGAGCCTGCTTGGGATTCTTTCTCT-T----CTCTGCCC

 174

Cf1 GCTCCTCTCCCTRCYYRTG(CT)1-15 (CT)2-8 CA--AA(TAAA)2 AA TC(T)n An

Cf2 GTGTCTCTGCCT-------(CT)1-5 (GT)1-18 (CT)2 CATGAA(TAAA)3 AA TC(T)n An

Cf3 GTGTCTCTGCCT-------(CT)2-10 (GT)1-5 CTAT CATGAA(TAAA)3 -A TC(T)n An

Mv GCCTCTCTGCCTRCTTGTG(CT)2-5 GT CA--AA(TAAA)3 -A TC(T)n An

Pv GCCYCTCTCCCTRCTTGTG(CT)4-14 GT (CT)2-6 CA--AA(TAAA)3 AA TC(T)n An

GgN14 ----CTCCTCCTGNTAGTG---------TG (CT)n CCCA--AA TAAA

Figure 2. Sequences of CanSINE families. Cf -*Canis familiaris*, Mv -*Mustela vision*, Pv -*Phoca vitulina*, Gg -*Gallus gallus*, lys tRNA: DK8101 - mouse, D50536 - mollusc*Loligo bleeekeri*.

Age estimates were obtained from analysis of subfamily members divergence from consensus by using methodology as described by Britten [8], Kapitonov and Jurka [9]. For estimation of actual level of nucleotide substitutions (K) Kimura's two-parameter model was used. The average age of subfamily divergence was estimated as T=K/0.002 [8, 10]. Result of age estimation is shown in Table 1.

Table 1. Age estimation for canSINE subfamilies

|  |  |  |  |
| --- | --- | --- | --- |
|   | **Divergence** | **Kimura's distance** | **Average time of divergence** |
| **cf1 subfamily** | **0.13** | **0.14** | **67** |
| **cf2 subfam** | **0.11** | **0.12** | **58** |
| **cf3 subfamily** | **0.05** | **0.05** | **24** |
| **Mink canSINE** | **0.08** | **0.09** | **43** |
| **Phoca canSINE** | **0.09** | **0.11** | **54** |
| **Gallus canSINE-like element** | **0.13** | **0.15** | **75** |

Age of dog cf1, mink and harbour seal subfamilies do not differ too much (67, 43, 54 million years). It can be suggested that propagation of canSINE started between the divergence of superfamilies Feloidea (approximately 55 million years ago) and the divergence of the family Canidae from the other canoid families (approximately 50 million years ago). This dating was obtained from analysis of fossils [11].

Presence of canSINE homologs in*Gallus*genome is an interesting evolutionary property of canSINE. Several explanations of this phenomenon can be suggested: 1)*Gallus*canSINE-like element is an artefact of experiments, 2) horizontal transfer of canSINE sequences, 3) canSINE-like*Gallus*sequences and canSINE independently originated from lys-tRNA, 4) canSINE progenitor was present in few copies before radiation of birds and mammals, although propagation of this element took place only in carnivore genomes and genomes of birds (at least in Gallus genome). The last hypothesis is probable since the homology between consensus sequences of canSINE and*Gallus*canSINE-like elements is not very high: Hc=85-90%. The third hypothesis is very questionable since homology between consensus sequences of canSINE and*Gallus*canSINE-like elements in tRNA-unrelated part in also highly significant. Although we cannot exclude the first and second hypothesis, we suggested that canSINE-like progenitor had been originated before radiation of birds and mammals. As results, canSINE could be suggested as a very good tracer of long-term evolution.

IBR was supported by the Russian Fund of Fundamental Research (grant N 96-04-49957).

**References**

1. W. Makalowski, "SINEs as a genomic scrap yard: an essay on genomic evolution", In "The Impact of Short Interspersed Elements (SINEs) on the Host Genome" (ed RJ Maraia), pp. 81-104, (R.G. Landes Company, Austin, 1995).
2. M.F. Minnick, L.C. Stilwell, J.M. Heineman, and G.L. Stiegler, "A highly repetitive DNA sequence possibly unique to Canids" Gene 110, 235 (1992).
3. M.V. Lavretieva, M.I. Rivkin, A.G. Shilov, M.L. Kobetz, I.B. Rogozin and O.L. Serov, "B2-like repetitive sequence from the X chromosome of the American mink (Mustela vison)" Mammalian Genome 1, 165 (1991).
4. D.W. Coltman and J.M. Wright, "Can SINEs: a family of tRNA-derived retroposons specific to the superfamily Canoidea" Nucleic Acids Res. 22, 2726 (1994).
5. S.F. Altschul, W. Gish, W. Miller, E.W. Myers and D.J. Lipman, "Basic local alignment search tool" J Mol Biol 215, 403 (1990).
6. W.R. Pearson and D.J. Lipman, "Improved tools for biological sequence comparison". Proc. Natl. Acad. Sci. USA 85, 2444 (1988).
7. D.G. Higgins and P.M. Sharp, "CLUSTAL: a package for performing multiple sequence alignments on a microcomputer" Gene 73, 237(1988).
8. R.J. Britten, "Evidence that most human Alu sequences were inserted in a process that ceased about 30 million years ago" Proc. Natl. Acad. Sci. USA 91, 6148 (1994).
9. V. Kapitonov and J. Jurka, "The age of Alu subfamilies" J. Mol.Evol. 42, 59 (1996).
10. W.-H. Li, M. Gouy, P.M. Sharp, C. O'hUigin and Y.-W. Yang, "Molecular phylogeny of Rodentia, Lagomorpha, Primates, Artiodactyla, and Carnivora and molecular clocks" Proc. Natl. Acad. Sci. USA 87, 6703 (1990).
11. J.L. Gittleman (ed.), (1989) Carnivore behaviour, ecology and evolution" (Cornell University Press, New York, 1989).