J Mol Evol (1993) 37:167–170

Journal of Molecular Evolution

A Model for the Mechanism of Initial Generation of Short Interspersed Elements (SINEs)

Norihiro Okada, Kazuhiko Ohshima

Faculty of Bioscience and Biotechnology, Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama 227, Japan

Abstract. Most animal genomes contain a large number of short interspersed elements (SINEs) that have a composite structure and contain a region that is homologous to a tRNA. The majority of SINEs have been found to be derived from a $tRNA^{Lys}$, being categorized as members of a superfamily of $tRNA^{Lys}$-related SINEs. The consensus sequences of five SINEs that belong to this superfamily were aligned. It was found that, in the tRNA-unrelated region, there are two sequence motifs that are almost identical among these five SINEs and are at a distance of 10–11 nucleotides from each other. This observation suggests a common evolutionary origin of these SINEs and/or some function(s) for these motifs. Similar sequences were unexpectedly found to be present in the sequences complementary to the U5 regions of several mammalian retroviruses whose primer is a $tRNA^{Lys}$. On the basis of these findings, we propose a possible model for the generation of SINEs whereby they are derived from a "strong stop DNA" with a primer tRNA that is an intermediate in the process of reverse transcription of certain retroviruses.

Key words: Short interspersed elements (SINEs) — Retroposon — $tRNA^{Lys}$ — Repetitive sequence — Strong stop DNA

Presented at the NATO Advanced Research Workshop on *Genome Organization and Evolution,* Spetsai, Greece, 16–22 September 1992

Correspondence to: N. Okada

In higher eukaryotes, protein-coding genes constitute 10% of the genome at most, the remainder of the genome being composed of a variety of repetitive sequences. Repetitive sequences in eukaryotic genomes can be classified into two groups: tandemly repeated sequences and dispersed sequences. The dispersed sequences can be further classified into two categories on the basis of size: long interspersed elements (LINEs), which include L1 sequences, and short interspersed elements (SINEs), such as the primate Alu family and the rodent type 1 and type 2 families (Singer 1982).

Repetitive sequences can be classified in a different way, according to the mechanism of their generation. Tandemly repeated sequences are generated by gene duplication at the DNA level (Singer and Berg 1991; Denison and Weiner 1982), while in the case of dispersed elements, another mechanism called retroposition has recently been characterized. During retroposition, information in cellular RNA can flow back into the genome via complementary DNA intermediates (Rogers 1985; Weiner et al. 1986).

Retroposon is the name given to a repetitive element that is amplified by retroposition. Nonviral retroposons can be classified into three groups, namely, processed retropseudogenes, SINEs and LINEs. In 1985, three laboratories including ours found that most mammalian SINEs, except for the human Alu (Weiner 1980; Ullu and Tschudi 1984), were derived from tRNAs or their genes (Lawrence et al. 1985; Daniels and Deininger 1985; Sakamoto and Okada 1985; Endoh and Okada 1986; Matsumoto et al. 1986). At that time, it was believed that SINEs are restricted to mammals, since none had

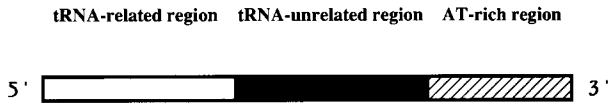

Fig. 1. The composite structure of a SINE.

Table 1. A $tRNA^{Lys}$-related SINE superfamily

SINE	Species	Reference
Galago type 2 family	*Galago*	Okada 1990
Rodent type 2 (B2) family	Mouse, rat, hamster	Sakamoto and Okada 1985
Tortoise *Pol*III SINE	Tortoise	Endoh et al. 1990
Charr *Fok*I family	*Salvelinus* spp.	Kido et al. 1991
Salmon *Sma*I family	Chum salmon, pink salmon	Kido et al. 1991
Squid SK family	Squid	Unpublished
Octopus OK family	Octopus	Unpublished
Tobacco TS family	Tobacco	Unpublished

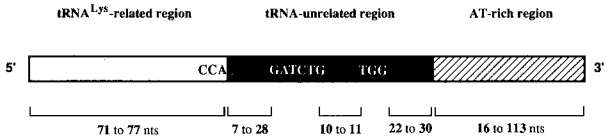

Fig. 2. A schematic representation of the location of the two conserved motifs in the tRNA-unrelated regions of SINEs.

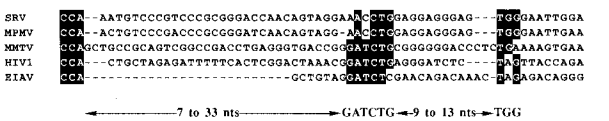

Fig. 3. A compilation of the sequences complementary to the U5 region adjacent to the primer-binding site (PBS) of several retroviruses. The 3′-terminal CCA of the primer $tRNA^{Lys}$ is indicated on the left side. SRV-1, MPMV, MMTV, HIV-1, and EIAV stand for simian acquired immune deficiency syndrome retrovirus, Mason-Pfizer monkey virus, mouse mammary tumor virus, human immunodeficiency virus type 1, and equine infectious anemia virus, respectively.

been found in the genomes of birds and *Drosophila* (Rogers 1985; Weiner et al. 1986). Recently, it was shown that SINEs are widespread in the animal kingdom (see, for recent reviews, Okada et al. 1991; Okada 1991a,b) and also in plants (Mochizuki et al. 1992; Yoshioka et al. unpublished results). These recent findings indicate that retroposition by way of SINE amplification continuously generates genetic and structural variations in the genomes of many more animal and plant species than had previously been supposed.

In general, SINEs are not simple tRNA pseudogenes but rather have a composite structure consisting of a region homologous to a tRNA, a tRNA-unrelated region, and an AT-rich region. The AT-rich region is sometimes not present or is replaced by an A-rich structure or some other repeating unit, such as $(TTG)_n$ (Yoshioka et al. unpublished results). The way in which the composite structure of SINEs is generated during evolution is unknown. In the present paper, we present a possible model for the generation of SINEs.

Results and Discussion

Figure 1 shows the composite structure of a SINE. Among a variety of tRNA species, $tRNA^{Lys}$ is the predominant progenitor of the tRNA-related region of SINEs (Okada et al. 1991; Okada 1991a,b). SINEs that are homologous to $tRNA^{Lys}$ are categorized as members of a superfamily of $tRNA^{Lys}$-related SINEs. Table 1 shows a compilation of SINEs that belong to this superfamily. It should be noted that distantly related species, such as rodents, squid, and tobacco, are included as hosts of this superfamily, suggesting that some unknown general mechanism may be involved in the generation of SINEs.

From among the members of this superfamily, we aligned the consensus sequences of five SINEs, namely, those of the rodent type 2 (B2) family (Krayev et al. 1982), tortoise *Pol*III SINE (Endoh et al. 1990), salmon *Sma*I family (Kido et al. 1991), charr *Fok*I family (Kido et al. 1991), and squid SK family (Ohshima et al. 1993). To our surprise, we found that, in the tRNA-unrelated regions of these SINEs, there are two sequence motifs that are almost identical in every case. These motifs are GATCTG and TGG, with a distance of 10–11 nucleotides between them. Figure 2 shows a schematic representation of the location of the two motifs in these five SINEs. Unexpectedly, sequences similar to these motifs were detected in the sequences complementary to the U5 regions of several retroviruses [SRV (Power et al. 1986), MPMV (Sonigo et al. 1986), MMTV (Fasel et al. 1983), HIV1 (Spire et al. 1989), EIAV (Derse et al. 1987)] and they were located at a distance from the primer-binding site similar to the distance between the 3′ end of tRNA-related region and these two motifs (Fig. 3). These retroviruses are known to use a $tRNA^{Lys}$ as a primer tRNA during reverse transcription.

On the basis of our findings, we present a possible model for the generation of SINEs, as shown in Fig. 4. The 3′-terminal sequence of a $tRNA^{Lys}$ (15–18 nts), including the CCA sequence, hybridizes to the primer-binding site (PBS) in the viral genome. Reverse transcription proceeds from the CCA end toward the 5′ end of the genome. The main product of reverse transcription in vitro is a single-stranded DNA with the $tRNA^{Lys}$ at its 5′ terminus, known as "strong stop DNA." During reverse transcription

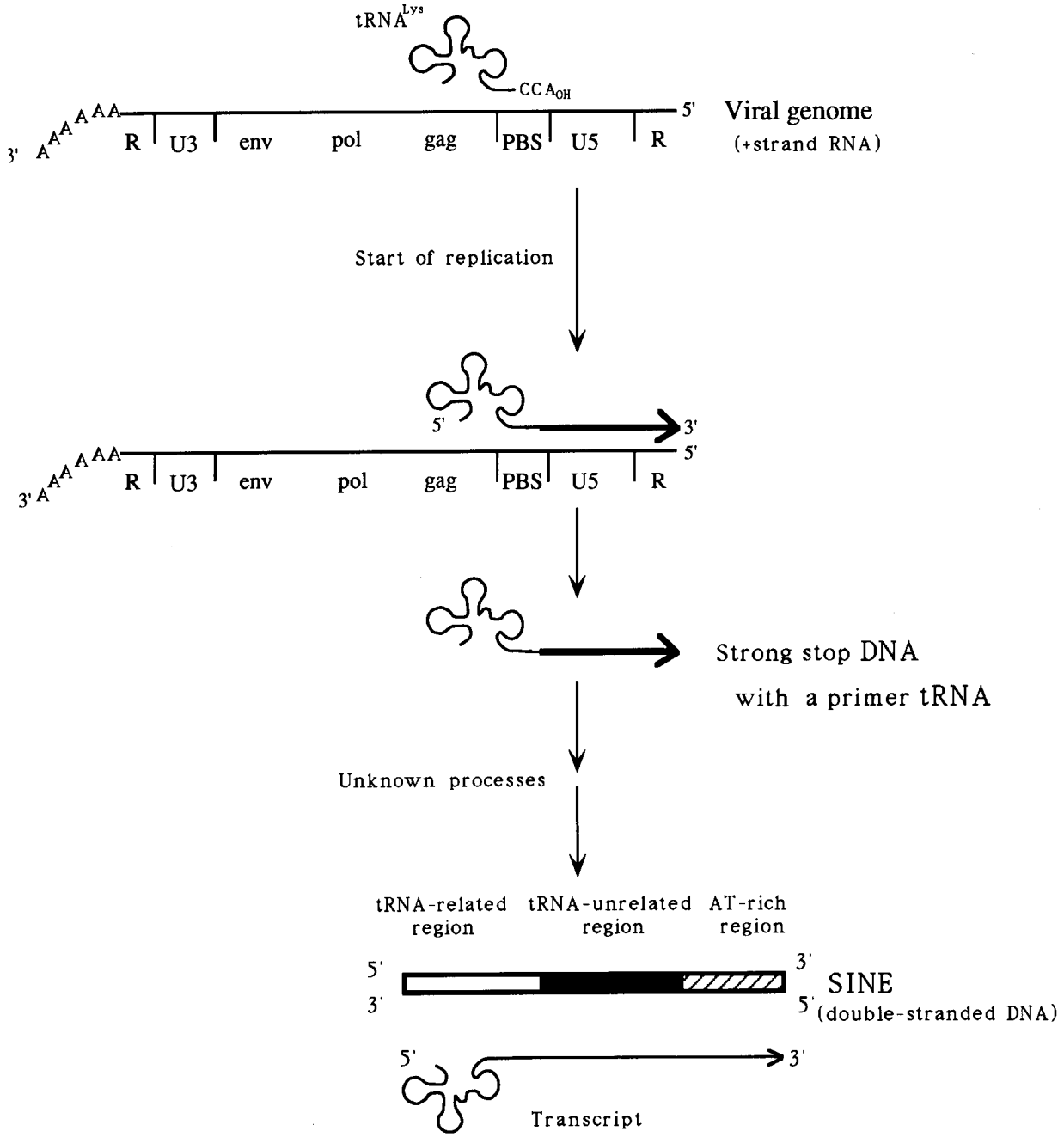

Fig. 4. A possible model for the generation of a SINE.

in vivo, the DNA "jumps" to the 3′ terminus of the viral genome due to the presence of the duplicated R region. From the sequence similarities described above, we propose a model wherein the strong stop DNA with the $tRNA^{Lys}$ becomes a SINE after several further unknown processes. The primer tRNA is not removed, and it is copied instead into DNA or inserted directly into the genome as a covalent tRNA-DNA hybrid, thereby creating a $tRNA^{Lys}$ pseudogene. A similar mechanism was originally suggested by Saigo (1986) and was later discussed several times by us (Okada 1991a; Endoh et al. 1990).

When the original model was proposed in 1986, it appeared to be very difficult to detect any similarities between SINEs and the U5 regions of retroviruses since the evolutionary rate of mutations in genes of retroviruses is one million times faster than that of nuclear genes (Gojobori and Yakoyama 1985). We suspected at that time that no similarities would remain, even if the model were correct. Sequence similarities between SINEs and retroviruses were discovered in the present study. However, they may not be significantly extensive to convince us of their evolutionary relationship. If we could isolate a pair of a new SINE and a new retrovirus in

which the sequences of the regions mentioned in this study were much more similar to each other, from a certain animal or plant, we would be more confident of the validity of our model.

With regard to the similarities between the five SINEs, the presence of the GATCTG and TGG sequences in the distantly related species convinces us that the similarities are significant. In this study, the similarities prompted us to speculate on the evolutionary origin of SINEs. Alternatively, however, it is possible that these motifs in SINEs have some function, having been generated independently in each species by convergent evolution.

References

Daniels GR, Deininger PL (1985) Repeat sequence families derived from mammalian tRNA genes. Nature (London) 317: 819–822

Denison RA, Weiner AM (1982) Human U1 RNA pseudogenes may be generated by both DNA- and RNA-mediated mechanisms. Mol Cell Biol 2:815–828

Derse D, Dorn PL, Levy L, Stephens RM, Rice NR, Casey JW (1987) Characterization of equine infectious anemia virus long terminal repeat. J Virol 61:743–747

Endoh H, Okada N (1986) Total DNA transcription in vitro: A procedure to detect highly repetitive and transcribable sequences with tRNA-like structures. Proc Natl Acad Sci USA 83:251–255

Endoh H, Nagahashi S, Okada N (1990) A highly repetitive and transcribable sequence in tortoise genome is probably a retroposon. Eur J Biochem 189:25–31

Fasel N, Buetti E, Firzlaff J, Pearson K, Diggelmann H (1983) Nucleotide sequence of the 5′ noncoding region and part of the *gag* gene of mouse mammary tumor virus: identification of the 5′ splicing site for subgenomic mRNAs. Nucleic Acids Res 11:6943–6955

Gojobori T, Yokoyama S (1985) Rates of evolution of the retroviral oncogene of Moloney murine sarcoma virus and of its cellular homologues. Proc Natl Acad Sci USA 82:4198–4201

Kido Y, Aono M, Yamaki T, Matsumoto K, Murata S, Saneyoshi M, Okada N (1991) Shaping and reshaping of salmonid genomes by amplification of tRNA-derived retroposons during evolution. Proc Natl Acad Sci USA 88:2326–2330

Krayev AS, Markusheva TV, Kramerov DA, Ryskov AP, Skryabin KG, Bayev AA, Georgiev GP (1982) Ubiquitous transposon-like repeats B1 and B2 of the mouse genome: B2 sequencing. Nucleic Acids Res 10:7461–7475

Lawrence CB, McDonnell DP, Ramsey WJ (1985) Analysis of repetitive sequence elements containing tRNA-like sequences. Nucleic Acids Res 13:4239–4252

Matsumoto K, Murakami K, Okada N (1986) Gene for lysine $tRNA_1$ may be a progenitor of the highly repetitive and transcribable sequences present in the salmon genome. Proc Natl Acad Sci USA 83:3156–3160

Mochizuki K, Umeda M, Ohtsubo H, Ohtsubo E (1992) Characterization of a plant SINE, p-SINE *1*, in rice genomes. Jpn J Genet 67:155–166

Ohshima K, Koishi R, Matsuo M, Okada N (1993) Several SINEs in distant species may have originated from a common ancestral retrovirus: characterization of a squid SINE and a possible mechanism for generation of tRNA-derived retroposons. Proc Natl Acad Sci USA in press

Okada N et al (1991) Evolution of repetitive sequences. In: Evolution of Life. Osawa S, Honjo T (eds) Springer-Verlag, Tokyo, pp 175–186

Okada N (1991a) SINEs: Short interspersed repeated elements of the eukaryotic genome. Trends Ecol Evol 6:358–361

Okada N (1991b) SINEs. Curr Opin in Genet Dev, 1:498–504

Okada N (1990) Transfer RNA-like structure of the human Alu family: implication of its generation mechanism and possible functions. J Mol Evol 31:500–510

Power MD Marx PA, Bryant ML, Gardner MB, Barr PJ, Luciw PA (1986) Nucleotide sequence of SRV-1, a type D simian aquired immune deficiency syndrome retrovirus. Science 231:1567–1572

Rogers J (1985) The structure and evolution of retroposons. Int Rev Cytol 93:231–279

Saigo K (1986) A *copia* primer pseudogen possibly generated by an abberant reverse transcription of a *copia*-related element in *Drosophila*. Nucleic Acids Res 14:7815

Sakamoto K, Okada N (1985) Rodent type 2 Alu family, rat identifier sequence, rabbit C family, and bovine or goat 73-bp repeat may have evolved from tRNA genes. J Mol Evol 22: 134–140

Singer MF (1982) SINEs and LINEs: Highly repeated short and long interspersed sequences in mammalian genomes. Cell 28: 433–434

Singer M, Berg P (1991) Genes and Genomes, University Science Books

Sonigo P, Barker C, Hunter E, Wain-Hobson S (1986) Nucleotide sequence of Mason-Pfizer monkey virus: an immunosuppressive D-type retrovirus. Cell 45:375–385

Spire B, Sire J, Zachar V, Rey F, Barré-Sinoussi F, Galibert F, Hampe A, Chermann J-C (1989) Nucleotide sequence of HIV1-NDK: a highly cytopathic strain of the human immunodeficiency virus. Gene 81:275–284

Ullu E, Tschudi C (1984) Alu sequences are processed 7SL RNA genes. Nature (London) 312:171–172

Weiner AM, Deininger PL, Efstratiadis A (1986) Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu Rev Biochem 55:631–661

Weiner AM (1980) An abundant cytoplasmic 7S RNA is complementary to the dominant interspersed middle repetitive DNA sequence family in the human genome. Cell 22:209–218