

SINEBase: a database and tool for SINE analysis

Nikita S. Vassetzky and Dmitri A. Kramerov*

Laboratory of Eukaryotic Genome Evolution, Engelhardt Institute of Molecular Biology, Moscow 119991, Russia

Received June 12, 2012; Revised October 11, 2012; Accepted November 5, 2012

ABSTRACT

SINEBase (<http://sines.eimb.ru>) integrates the revisited body of knowledge about short interspersed elements (SINEs). A set of formal definitions concerning SINEs was introduced. All available sequence data were screened through these definitions and the genetic elements misidentified as SINEs were discarded. As a result, 175 SINE families have been recognized in animals, flowering plants and green algae. These families were classified by the modular structure of their nucleotide sequences and the frequencies of different patterns were evaluated. These data formed the basis for the database of SINEs. The SINEBase website can be used in two ways: first, to explore the database of SINE families, and second, to analyse candidate SINE sequences using specifically developed tools. This article presents an overview of the database and the process of SINE identification and analysis.

INTRODUCTION

Short interspersed elements (SINEs) are mobile genetic elements invading genomes of most higher eukaryotes (exceeding 10% of some genomes). Although these genomic parasites can be deleterious to the cell, the long-term being in the genome has made SINEs a valuable factor of genetic variation, providing regulatory elements for gene expression, alternative splice sites, polyadenylation signals and even functional RNA genes (1–4). At the same time, the system and nomenclature of SINEs remain to a large extent unarticulated. SINEBase is a manually curated database of SINE families known to date. It aims to be a resource for scientists working on mobile elements as well as for a wide range of biologists analysing nucleic acid sequences. SINEBase can be considered as a compendium of SINEs; its toolset allows individual SINE sequences to be attributed to known SINE families and/or analysed.

Definitions

Retro(trans)posons are genetic elements that can amplify themselves in eukaryotic genomes, which requires an

RNA intermediate, and thus, transcription and reverse transcription. Retrotransposons are divided into three classes: long terminal repeat (LTR) elements, long interspersed elements (LINEs) and SINEs. The elements that encode the enzyme activities, providing for the reverse transcription and integration of the DNA copy into the genome, are called *autonomous* transposons. *Nonautonomous* retrotransposons rely on the enzyme machinery of autonomous transposons. LTR retrotransposons and LINEs can be autonomous or nonautonomous; and their genomic copies are transcribed by the cellular RNA polymerase II (5,6).

SINEs are defined as relatively short (<700 bp) nonautonomous retrotransposons transcribed by the cellular RNA polymerase III (pol III) from an internal promoter, whereas their reverse transcription depends on the reverse transcriptase of partner LINEs. Eukaryotic genomes can harbor hundreds thousands (sometimes more) of SINE copies; copies originating from a common ancestral SINE can differ from each other by single-nucleotide alterations as well as by longer internal deletions or duplications (SINEs with such duplication are called *quasidimeric*). Some of them can become founders of new SINE *subfamilies*.

SINEs consist of two or more *modules*; typically, *head*, *body* and *tail*. The 5'-terminal head originates from one of cellular RNAs synthesized by pol III: tRNA, 7SL RNA or 5S rRNA. The origin of the body is either unknown or it descends from a partner LINE. SINEs with such a region mimic LINE RNA in the reverse transcription (7). The body can also contain a central domain shared by distant SINE families (CORE and similar domains). The 3'-terminal tail is a sequence of variable length consisting of simple (often degenerate) repeats. In addition, two SINEs can combine into a *dimeric* SINE, thus, giving rise to a new SINE family. SINEs consisting of the head and tail only are called *simple*, whereas dimeric, trimeric, etc. are *complex* SINEs. Various aspects of SINE structure, biology and evolution have been reviewed elsewhere (4,8,9).

We consider SINEs as (i) short (<1 kb) interspersed (nontandem) genomic repeats; (ii) present in at least 100 copies per genome (except certain genomes where repetitive elements are not abundant, e.g. *Arabidopsis thaliana*); (iii) with at least 60% identity with a tRNA species (10),

*To whom correspondence should be addressed. Tel: +7 499 135 9997; Fax: +7 499 135 1405; Email: kramerov@eimb.ru

5S rRNA or 7SL RNA in at least 60-nt overlap (with a few exceptions where the element transcription by pol III was confirmed experimentally). We found that pol III promoters (e.g. boxes A and B) can serve only as an indication (but not a proof) that the sequence belongs to SINEs. (Even when more sophisticated methods of pol III promoter identification, e.g. with position frequency matrices, were used, the proportion of false positives and/or misses remained high for different 'stringency' values). SINEs should be distinguished from RNA pseudogenes: the pseudogenes are generated by the reverse transcription of the functional RNAs of cellular origin (e.g. 5S rRNA) rather than of SINE RNAs transcribed from their genomic copies. In practical terms, most SINEs have extra (body) sequences, whereas simple SINEs have characteristic substitutions/indels shared with their source gene but not with the cellular RNA gene. In addition, SINEs significantly outnumber RNA pseudogenes.

The notion of 'SINE family' is widely used but not clearly defined. We consider SINE family as a set of SINEs (i) of a common origin and (ii) consisting of the same modules (except the tail, which can vary even in the same species). Thus, similar SINEs with different LINE-derived regions (e.g. mammalian Ther-2 and Mar-1) belong to different families. Long insertions are considered as modules. At the same time, internal deletions or duplications within modules do not give birth to a new family; although a combination of complete or almost complete SINEs (complex SINEs) is considered as a new family (thus, pB1 and quasidimeric B1 are subfamilies of the same family, whereas dimeric Alu represents a distinct family). Finally, there are a few SINEs with similar structure but of independent origin (e.g. simple SINEs: ID in rodents, vic-1 in camels and DAS-I in armadillos), thus, considered as different families.

DATABASE OF SINE FAMILIES

Data acquisition

We extracted consensus sequences of SINE families largely from two sources, original publications and the Repbase Update (RU; ver. 16.07) database (11). In many cases, they were refined in the available sequence databases. The consensus sequences were compared with the sequences of other SINEs, LINEs, tRNA species, 5S rRNA and 7SL RNA to identify their modules. Similar sequences were aligned and the differences were analysed. The elements composed of the same modules were considered as the same SINE family. There were particularly knotty cases, e.g. the CHRS family. This SINE is a quasioligomer, it contains a ~20-nt degenerate motif, which can be tandemly repeated more than 10 times. The variants differing in the number of these repeats were previously recognized as different SINE families (CHRS, CHR-2, CHRL, etc.). Multiple alignments clarifying such cases can be found in Supplementary Alignments S1. As a result, 175 SINE families were recognized according to the above definitions.

SINEBase organization

The heart of the database is the *SINETable* (also available as Supplementary Table S1) visualizing the main data about all SINE families known to date (length, distribution, copy number, schematic structure, etc.). The table contents can be limited to certain taxa and sorted by some characters (e.g. tail sequence). It contains links to SINE family-specific data (e.g. consensus sequence or publications) or to term descriptions. The databases of consensus sequences of SINE families, central domains and LINE-derived regions can be downloaded in the *Download* section, whereas individual consensus sequences and the multiple alignments are accessible as Supplementary Sequences S1 and Supplementary Alignments S2 and S3, respectively.

SINEBase tool

Based on our long-term experience in SINE analysis, we offer a toolset for the identification of SINE families and modules (*SINESearch*). This tool can also ascertain that the sequence of interest is not a SINE or that it belongs to an unknown SINE family. In the latter case, *SINESearch* can be used to analyse the modules of a new SINE.

It is a FASTA-based search that uses parameters other than the FASTA's statistical significance test to select sequences. This obviates two limitations of FASTA (as well as BLAST etc.) in the case of relatively short and degenerate similarities between nucleotide sequences: a bias to short (almost) perfect matches, whereas the goal is full-length and significant similarities; and missing significant hits when the bank includes many sequences similar to query. The search banks include our collections of full-length SINEs and their modules (certain RNAs, central domains and LINE-derived regions).

SINESearch is simple to use and fast. The search parameters used, *overlap length* and *sequence identity*, are biologically sensible and allow easy adjustment of hit selection. Query sequence can be manually input or uploaded. *SINESearch* offers four banks: *SINEBank* (consensus sequences of SINE families), *RNABank* (human tRNA species (10) plus 7SL RNA and 5S rRNA), *LINEBank* (SINE consensus sequences derived from partner LINEs) and *COREBank* [consensus sequences of central (CORE, Deu-, V-, Ceph-, α - and β -) domains].

The recommended protocol for the analysis of putative SINE sequences (explained in detail in the *Help* section) includes the following steps: preliminary analysis of a sequence of interest to exclude non-SINE sequences; *SINESearch* against the *SINEBank* to identify SINEs that belong to known SINE families and *SINESearch* against other banks to identify individual modules of a putative SINE.

SINE data analysis

The length of SINE consensus sequences without tail ranges from 75 to 662 nt, with the mean length of 253 nt (Figure 1). In terms of structure (Figure 2), the majority of families are monomeric (87%) tRNA-derived (84%; green

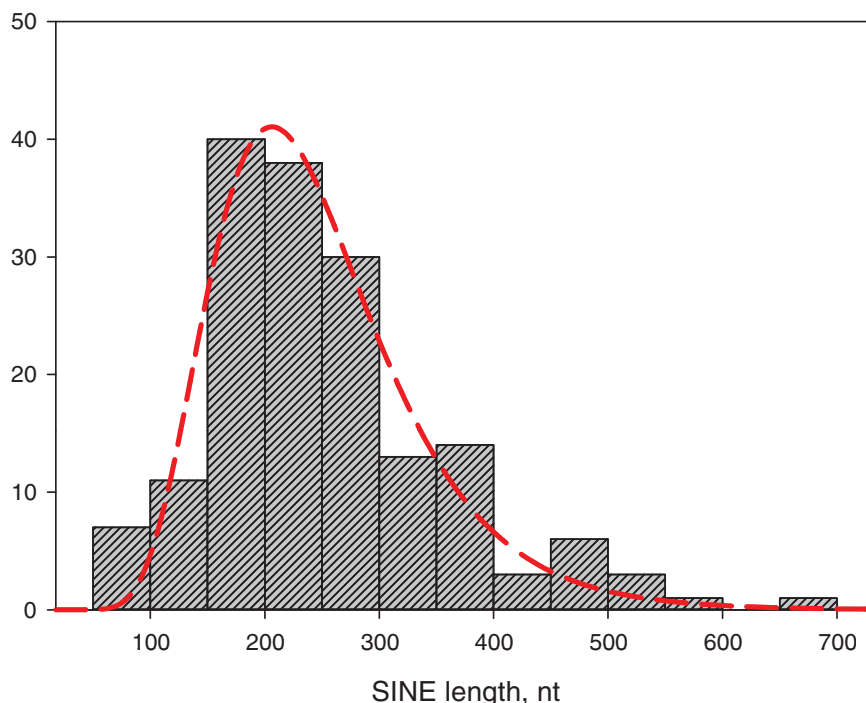


Figure 1. Length distribution for 175 eukaryotic SINE families (without tail). The range is from 75 to 662 nt with the mean and median length of 253 and 236 nt, respectively.

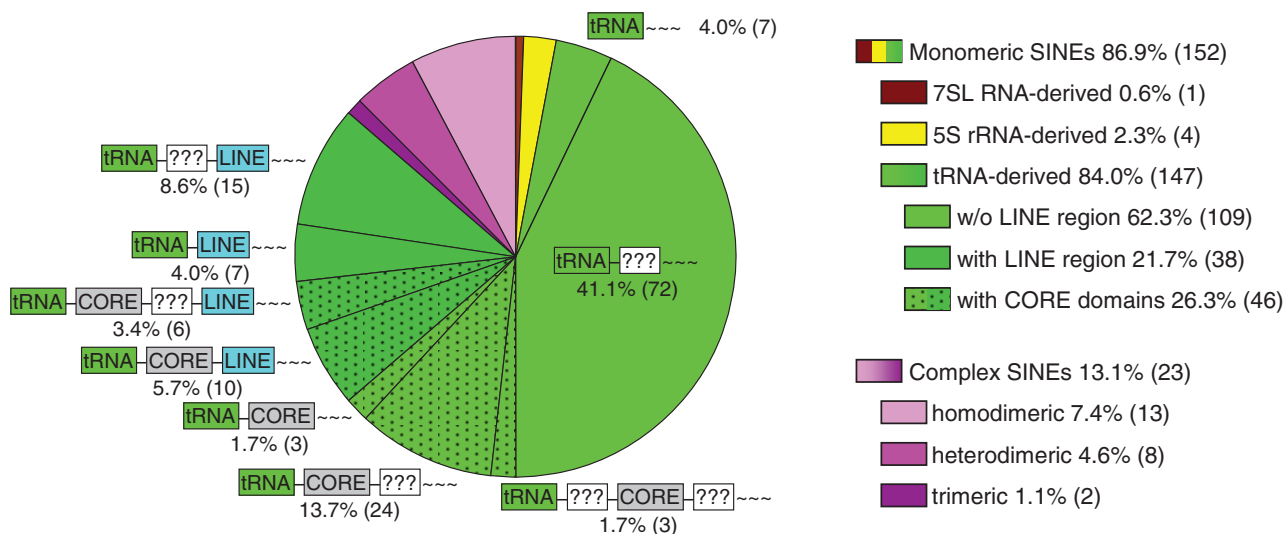


Figure 2. Occurrence of different SINE structures. Complex SINEs are highlighted by shades of purple; brown and yellow sectors represent 7SL RNA- and 5S rRNA-derived SINEs, respectively; tRNA-derived are shown in shades of green (dark and light hues correspond to SINEs with and without LINE-derived region, respectively). Dotted sectors indicate SINEs with CORE domains; schematic SINE structures (as in SINETable/Supplementary Table S1) are shown next to or over the corresponding sectors. Percentage is the fraction of a structure among 175 SINE families, and the number in parentheses is the number of SINE families representing the structure.

sectors in Figure 2) SINEs. There are roughly 3 times less SINE families with the LINE-derived region (dark green sectors) than without it (light green sectors), although this ratio can decrease as new partner LINES become identified. More than a quarter families contain CORE and similar domains (dotted sectors). The most common SINE structure is a tRNA-derived head followed by a body of unknown origin and a tail (41%); other patterns

range from 2 to 14%. Complex SINE families amount to 13% (purple sectors in Figure 2).

The collection of consensus sequences was further analysed in an attempt to identify similar patterns in their structure. All tRNA-derived sequences of SINE families were used to generate a sequence logo, which was compared with that of human tRNA genes (Figure 3; Supplementary Alignments S4). Overall, the same

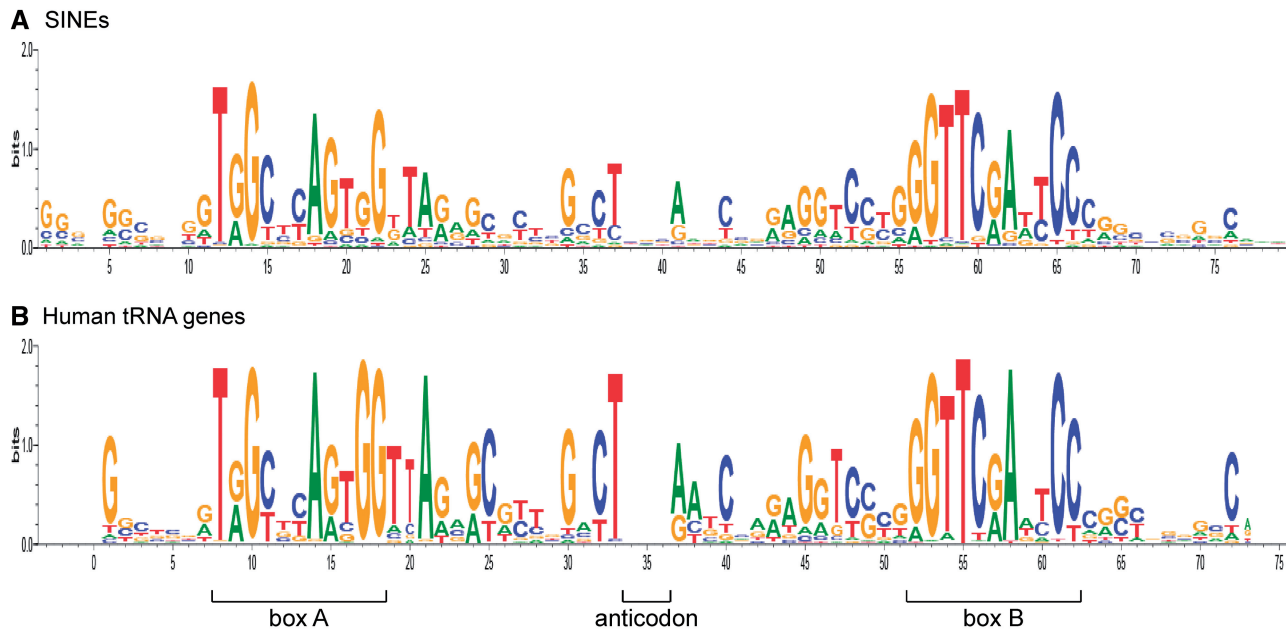


Figure 3. Conservation of tRNA-derived sequences in SINEs. Sequence logos of (A) 175 tRNA-derived sequences in SINEs (including those in the second and third monomers of complex SINEs) and (B) 359 human tRNA genes (10) were generated by WebLogo 3.1 (13). The multiple alignments were edited to eliminate gaps in the logos. The original alignments of tRNAs and tRNA-derived sequences in SINEs are available in Supplementary Alignments S4.

sequence pattern is observed in both cases, although it is less pronounced for SINEs (i.e. SINE sequences are more variable). SINEs have a short G-rich extra sequence at the 5'-end compared with tRNAs (and of course extra downstream sequences).

Special surveys were carried out for the body region targeted at the central CORE-like domains and LINE-derived regions. This allowed us to identify such domains in certain SINEs, where they remained unnoticed (e.g. the CORE domain was found in several sea urchin SINEs), as well as to identify two new central domains named α and β . As a result, the new consensus sequences were generated for most central domains (CORE, Deu, V, Ceph, α and β) (Supplementary Alignments S2). A similar analysis of the body 3' terminal regions allowed us to generate multiple alignments and consensus sequences for four LINE-derived regions corresponding to Bov-B, CR1 and two L2 LINES (Supplementary Alignments S3).

Similar resources

The most comprehensive up-to-date database of repetitive genomic elements (REs) is RU (11). *De facto*, it has become the standard source for RE research and nomenclature. RU includes many other types of REs apart from SINEs, whereas SINE consensus sequences represent families and subfamilies from groups of organisms and individual genomes in the same pool. Clearly, SINEBase that covers only SINE families cannot be considered as a RU replacement. At the same time, our analysis revealed a number of discrepancies between SINEBase and RU, what we believe stem from certain errors and ambiguities in RU (Supplementary Table S2). (i) As many as 80 records annotated in RU as SINEs (527 by the analysis

time) were not included in SINEBase as they correspond to other RE types (largely, LINES); (ii) SINEBase assigns consistent names to the same SINEs in different species and to SINE subfamilies, which in RU can be assigned different names (only 130 RU records correspond to SINEBase families, whereas 258 correspond to subfamilies and species variants); (iii) A substantial fraction of SINEBase families (45 in total) are missing from RU; and (iv) Finally, SINEBase uses a straightforward SINE nomenclature, which in most cases relies on the previously described SINEs, whereas RU tends to rename them. As an example, the RU naming scheme (39) (which changed several times) includes 43 records starting from 'SINE2-' and a text with numerous SINE2-2_CQ, SINE2-2_NV and SINE2-2_SP is hard to read. We renamed such families by omitting the redundant 'SINEx'. Supplementary Table S2 lists all RU records annotated as SINEs and describes their status in SINEBase.

Furthermore, the RepeatMasker program (<http://www.repeatmasker.org>) routinely used to identify SINEs relies on (slightly modified) RU records. RepeatMasker finds the best hit among RU sequences based on certain statistical parameters, such that a high similarity over short fragments can be considered more significant than a lower similarity throughout the element; this is particularly true for short sequences. At the same time, different SINE families can share the same (e.g. 5S rRNA-derived) module, and the sequences in this region can be highly similar, whereas those in the other regions are dissimilar. The situation is even worse for SINE subfamilies distinguished by diagnostic characters (often single nucleotide), as RepeatMasker considers them on par with random (non-diagnostic) mutations. Because many RU records

belong to the same SINE family, the RepeatMasker can recognize a set of sequences of the same family as several different SINEs.

As a result, blind reliance on this tool often leads to confusing misidentifications of SINEs or even anecdotal errors in otherwise competent publications, e.g. hundreds thousands of Alu copies in the genomes of mouse and rat (rodents have a different much shorter B1 SINE; Alu is limited to primates) or a single 51-nt B2 in the human genome (SINEs are longer and repetitive; most likely it is a tRNA pseudogene) (12). We have designed the SINEsearch tool to be free from these limitations.

SINEBase website

SINEBase is hosted on an Apache web server using CGI-Perl and JavaScript to generate dynamic HTML pages. It is fully functional with major web browsers. Some older browsers tested still allow the backbone functions of the site, whereas some decorations may not work (e.g. *SINETable* sorting). The database will be updated as new SINE data become available to us (at least biannually). We encourage the submission of new data on SINE families. SINEBase is freely available at <http://sines.eimb.ru>. There are no access restrictions for academic and commercial use. We kindly ask all users to cite this article if they use SINEBase in their publications.

CONCLUSION

As more and more genome sequences become available, the number of known SINEs will grow and new researchers will be involved in their analysis. SINEBase is aimed to bring some order to the system of SINEs and to set a basis for further studies on these genomic elements. The database of SINE consensus sequences and motifs will be updated as new SINEs are described. We will develop new tools to assist in SINE analysis (the identification of TSDs and internal duplications are first in the list). We appreciate feedback from SINEBase users to improve the service.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Caption 1, Supplementary Tables 1 and 2, Supplementary Sequence 1, Supplementary Alignments 1–4 and Supplementary References [13–99].

ACKNOWLEDGMENTS

Thanks are due to Dr. Yuri Kravatsky for his expert advice on web programming. All SINE elements originally published in Repbase have been reproduced with permission from the Genetic Information Research Institute, which reserves all rights in the material. We thank Drs. J. Jurka and V. V. Kapitonov for allowing access to their data.

FUNDING

Molecular and Cellular Biology Program of the Russian Academy of Sciences; Russian Foundation for Basic Research [project nos. 13-04-01678-a and 11-04-00439-a]. Funding for open access charge: Personal funds.

Conflict of interest statement. None declared.

REFERENCES

- Makalowski, W. (2000) Genomic scrap yard: how genomes utilize all that junk. *Gene*, **259**, 61–67.
- Belancio, V.P., Hedges, D.J. and Deininger, P. (2008) Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. *Genome Res.*, **18**, 343–358.
- Okada, N., Sasaki, T., Shimogori, T. and Nishihara, H. (2010) Emergence of mammals by emergency: exaptation. *Genes Cells*, **15**, 801–812.
- Kramerov, D.A. and Vassetzky, N.S. (2011) SINEs. Wiley interdisciplinary reviews. *RNA*, **2**, 772–786.
- Deininger, P.L. and Batzer, M.A. (2002) Mammalian retroelements. *Genome Res.*, **12**, 1455–1465.
- Gogvadze, E. and Buzdin, A. (2009) Retroelements and their impact on genome evolution and functioning. *Cell. Mol. Life Sci.*, **66**, 3727–3742.
- Okada, N., Hamada, M., Ogiwara, I. and Ohshima, K. (1997) SINEs and LINEs share common 3' sequences: a review. *Gene*, **205**, 229–243.
- Ohshima, K. and Okada, N. (2005) SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. *Cytogenet. Genome Res.*, **110**, 475–490.
- Kramerov, D.A. and Vassetzky, N.S. (2011) Origin and evolution of SINEs in eukaryotic genomes. *Heredity*, **107**, 487–495.
- Juhling, F., Morl, M., Hartmann, R.K., Sprinzl, M., Stadler, P.F. and Putz, J. (2009) tRNADB 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.*, **37**, D159–D162.
- Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, **110**, 462–467.
- Naked Mole Rat Genome Sequencing Consortium. (2011) Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature*, **479**, 223–227.
- Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
- Piskurek, O., Nishihara, H. and Okada, N. (2009) The evolution of two partner LINE/SINE families and a full-length chromodomain-containing Ty3/Gypsy LTR element in the first reptilian genome of *Anolis carolinensis*. *Gene*, **441**, 111–118.
- Ogiwara, I., Miya, M., Ohshima, K. and Okada, N. (2002) V-SINEs: a new superfamily of vertebrate SINEs that are widespread in vertebrate genomes and retain a strongly conserved segment within each repetitive unit. *Genome Res.*, **12**, 316–324.
- Takahashi, K., Terai, Y., Nishida, M. and Okada, N. (1998) A novel family of short interspersed repetitive elements (SINEs) from cichlids: the patterns of insertion of SINEs at orthologous loci support the proposed monophyly of four major groups of cichlid fishes in Lake Tanganyika. *Mol. Biol. Evol.*, **15**, 391–407.
- Zhao, F., Qi, J. and Schuster, S.C. (2009) Tracking the past: interspersed repeats in an extinct Afrotherian mammal, *Mammuthus primigenius*. *Genome Res.*, **19**, 1384–1392.
- Nikaido, M., Nishihara, H., Hukamoto, Y. and Okada, N. (2003) Ancient SINEs from African Endemic Mammals. *Mol. Biol. Evol.*, **20**, 522–527.
- Deininger, P.L., Jolly, D.J., Rubin, C.M., Friedmann, T. and Schmid, C.W. (1981) Base sequence studies of 300 nucleotide renatured repeated human DNA clones. *J. Mol. Biol.*, **151**, 17–33.
- Yasui, Y., Nasuda, S., Matsuoka, Y. and Kawahara, T. (2001) The Au family, a novel short interspersed element (SINE) from *Aegilops umbellulata*. *Theor. Appl. Genet.*, **102**, 463–470.

21. Fawcett, J.A., Kawahara, T., Watanabe, H. and Yasui, Y. (2006) A SINE family widely distributed in the plant kingdom and its evolutionary history. *Plant Mol. Biol.*, **61**, 505–514.
22. Kido, Y., Himberg, M., Takasaki, N. and Okada, N. (1994) Amplification of distinct subfamilies of short interspersed elements during evolution of the Salmonidae. *J. Mol. Biol.*, **241**, 633–644.
23. Matveev, V. and Okada, N. (2009) Retroposons of salmonoid fishes (Actinopterygii: Salmonoidei) and their evolution. *Gene*, **434**, 16–28.
24. Krayev, A.S., Kramerov, D.A., Skryabin, K.G., Ryskov, A.P., Bayev, A.A. and Georgiev, G.P. (1980) The nucleotide sequence of the ubiquitous repetitive DNA sequence B1 complementary to the most abundant class of mouse fold-back RNA. *Nucleic Acids Res.*, **8**, 1201–1215.
25. Veniaminova, N.A., Vassetzky, N.S. and Kramerov, D.A. (2007) B1 SINEs in different rodent families. *Genomics*, **89**, 678–686.
26. Kramerov, D.A. and Vassetzky, N.S. (2001) Structure and origin of a novel dimeric retroposon B1-dID. *J. Mol. Evol.*, **52**, 137–143.
27. Krayev, A.S., Markusheva, T.V., Kramerov, D.A., Ryskov, A.P., Skryabin, K.G., Bayev, A.A. and Georgiev, G.P. (1982) Ubiquitous transposon-like repeats B1 and B2 of the mouse genome: B2 sequencing. *Nucleic Acids Res.*, **10**, 7461–7475.
28. Lee, I.Y., Westaway, D., Smit, A.F., Wang, K., Seto, J., Chen, L., Acharya, C., Ankener, M., Baskin, D., Cooper, C. et al. (1998) Complete genomic sequence and analysis of the prion protein gene region from three mammalian species. *Genome Res.*, **8**, 1022–1037.
29. Nishihara, H., Smit, A.F. and Okada, N. (2006) Functional noncoding sequences derived from SINEs in the mammalian genome. *Genome Res.*, **16**, 864–874.
30. Adams, D.S., Eickbush, T.H., Herrera, R.J. and Lizardi, P.M. (1986) A highly reiterated family of transcribed oligo(A)-terminated, interspersed DNA elements in the genome of *Bombyx mori*. *J. Mol. Biol.*, **187**, 465–478.
31. Xu, J., Liu, T., Li, D., Zhang, Z., Xia, Q. and Zhou, Z. (2010) BmSE, a SINE family with 3' ends of (ATTT) repeats in domesticated silkworm (*Bombyx mori*). *J. Genet. Genomics*, **37**, 125–135.
32. Lenstra, J.A., van Boxtel, J.A., Zwaagstra, K.A. and Schwerin, M. (1993) Short interspersed nuclear element (SINE) sequences of the Bovidae. *Anim. Genet.*, **24**, 33–39.
33. Cheng, J.F., Printz, R., Callaghan, T., Shuey, D. and Hardison, R.C. (1984) The rabbit C family of short, interspersed repeats. Nucleotide sequence determination and transcriptional analysis. *J. Mol. Biol.*, **176**, 1–20.
34. Krane, D.E., Clark, A.G., Cheng, J.F. and Hardison, R.C. (1991) Subfamily relationships and clustering of rabbit C repeats. *Mol. Biol. Evol.*, **8**, 1–30.
35. Lavrent'eva, M.V., Rivkin, M.I., Shilov, A.G., Kobets, M.L., Rogozin, I.B. and Serov, O.L. (1989) B2-like repetitive sequence in the genome of the American mink. *Dokl. Akad. Nauk SSSR*, **307**, 226–228.
36. Vassetzky, N.S. and Kramerov, D.A. (2002) CAN—a pan-carnivore SINE family. *Mamm. Genome*, **13**, 50–57.
37. Shimamura, M., Abe, H., Nikaido, M., Ohshima, K. and Okada, N. (1999) Genealogy of families of SINEs in cetaceans and artiodactyls: the presence of a huge superfamily of tRNA(Glu)-derived families of SINEs. *Mol. Biol. Evol.*, **16**, 1046–1060.
38. Nikaido, M., Matsuno, F., Abe, H., Shimamura, M., Hamilton, H., Matsubayashi, H. and Okada, N. (2001) Evolution of CHR-2 SINEs in cetartiodactyl genomes: possible evidence for the monophyletic origin of toothed whales. *Mamm. Genome*, **12**, 909–915.
39. Kapitonov, V.V. and Jurka, J. (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genet.*, **9**, 411–412.
40. Simmen, M.W. and Bird, A. (2000) Sequence analysis of transposable elements in the sea squirt, *Ciona intestinalis*. *Mol. Biol. Evol.*, **17**, 1685–1694.
41. Endoh, H., Nagahashi, S. and Okada, N. (1990) A highly repetitive and transcribable sequence in the tortoise genome is probably a retroposon. *Eur. J. Biochem.*, **189**, 25–31.
42. Sasaki, T., Takahashi, K., Nikaido, M., Miura, S., Yasukawa, Y. and Okada, N. (2004) First application of the SINE (short interspersed repetitive element) method to infer phylogenetic relationships in reptiles: an example from the turtle superfamily Testudinoidea. *Mol. Biol. Evol.*, **21**, 705–715.
43. Piskurek, O., Nikaido, M., Boeadi, B., Baba, M. and Okada, N. (2003) Unique mammalian tRNA-derived repetitive elements in dermopterans: the t-SINE family and its retrotransposition through multiple sources. *Mol. Biol. Evol.*, **20**, 1659–1668.
44. Schmitz, J. and Zischler, H. (2003) A novel family of tRNA-derived SINEs in the colugo and two new retrotransposable markers separating dermopterans from primates. *Mol. Phylogenet. Evol.*, **28**, 341–349.
45. Izsvak, Z., Ivics, Z., Garcia-Estefania, D., Fahrenkrug, S.C. and Hackett, P.B. (1996) DANA elements: a family of composite, tRNA-derived short interspersed DNA elements associated with mutational activities in zebrafish (*Danio rerio*). *Proc. Natl Acad. Sci. USA*, **93**, 1077–1081.
46. Borodulina, O.R. and Kramerov, D.A. (2005) PCR-based approach to SINE isolation: simple and complex SINEs. *Gene*, **349**, 197–205.
47. Churakov, G., Smit, A.F., Brosius, J. and Schmitz, J. (2005) A novel abundant family of retroposed elements (DAS-SINEs) in the nine-banded armadillo (*Dasypus novemcinctus*). *Mol. Biol. Evol.*, **22**, 886–893.
48. Serdobova, I.M. and Kramerov, D.A. (1998) Short retroposons of the B2 superfamily: evolution and application for the study of rodent phylogeny. *J. Mol. Evol.*, **46**, 202–214.
49. Sakagami, M., Ohshima, K., Mukoyama, H., Yasue, H. and Okada, N. (1994) A novel tRNA species as an origin of short interspersed repetitive elements (SINEs). Equine SINEs may have originated from tRNA(Ser). *J. Mol. Biol.*, **239**, 731–735.
50. Borodulina, O.R. and Kramerov, D.A. (2001) Short interspersed elements (SINEs) from insectivores. Two classes of mammalian SINEs distinguished by A-rich tail structure. *Mamm. Genome*, **12**, 779–786.
51. Tu, Z. (1999) Genomic and evolutionary analysis of Feilai, a diverse family of highly reiterated SINEs in the yellow fever mosquito, *Aedes aegypti*. *Mol. Biol. Evol.*, **16**, 760–772.
52. Tu, Z., Li, S. and Mao, C. (2004) The changing tails of a novel short interspersed element in *Aedes aegypti*: genomic evidence for slippage retrotransposition and the relationship between 3' tandem repeats and the poly(dA) tail. *Genetics*, **168**, 2037–2047.
53. Ogiwara, I., Miya, M., Ohshima, K. and Okada, N. (1999) Retropositional parasitism of SINEs on LINEs: identification of SINEs and LINEs in elasmobranchs. *Mol. Biol. Evol.*, **16**, 1238–1250.
54. Kido, Y., Aono, M., Yamaki, T., Matsumoto, K., Murata, S., Saneyoshi, M. and 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.
55. Milner, R.J., Bloom, F.E., Lai, C., Lerner, R.A. and Sutcliffe, J.G. (1984) Brain-specific genes have identifier sequences in their introns. *Proc. Natl Acad. Sci. USA*, **81**, 713–717.
56. Akasaki, T., Nikaido, M., Nishihara, H., Tsuchiya, K., Segawa, S. and Okada, N. (2010) Characterization of a novel SINE superfamily from invertebrates: "Ceph-SINEs" from the genomes of squids and cuttlefish. *Gene*, **454**, 8–19.
57. Gogolevsky, K.P. and Kramerov, D.A. (2006) Short interspersed elements (SINEs) of the *Geomyoidea superfamily* rodents. *Gene*, **373**, 67–74.
58. Churakov, G., Sadasivuni, M.K., Rosenbloom, K.R., Huchon, D., Brosius, J. and Schmitz, J. (2010) Rodent evolution: back to the root. *Mol. Biol. Evol.*, **27**, 1315–1326.
59. Bejerano, G., Lowe, C.B., Ahituv, N., King, B., Siepel, A., Salama, S.R., Rubin, E.M., Kent, W.J. and Haussler, D. (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature*, **441**, 87–90.
60. Bradfield, J.Y., Locke, J. and Wyatt, G.R. (1985) An ubiquitous interspersed DNA sequence family in an insect. *DNA*, **4**, 357–363.
61. Munemasa, M., Nikaido, M., Nishihara, H., Donnellan, S., Austin, C.C. and Okada, N. (2008) Newly discovered young CORE-SINEs in marsupial genomes. *Gene*, **407**, 176–185.
62. Warren, W.C., Hillier, L.W., Marshall Graves, J.A., Birney, E., Ponting, C.P., Grutzner, F., Belov, K., Miller, W., Clarke, L., Chinwalla, A.T. et al. (2008) Genome analysis of the

- platypus reveals unique signatures of evolution. *Nature*, **453**, 175–183.
63. Gilbert, N. and Labuda, D. (2000) Evolutionary inventions and continuity of CORE-SINEs in mammals. *J. Mol. Biol.*, **298**, 365–377.
 64. Gentles, A.J., Wakefield, M.J., Kohany, O., Gu, W., Batzer, M.A., Pollock, D.D. and Jurka, J. (2007) Evolutionary dynamics of transposable elements in the short-tailed opossum *Monodelphis domestica*. *Genome Res.*, **17**, 992–1004.
 65. Gogolevsky, K.P., Vassetzky, N.S. and Kramerov, D.A. (2009) 5S rRNA-derived and tRNA-derived SINEs in fruit bats. *Genomics*, **93**, 494–500.
 66. Nishihara, H., Kuno, S., Nikaido, M. and Okada, N. (2007) MyrSINEs: a novel SINE family in the anteater genomes. *Gene*, **400**, 98–103.
 67. Ohshima, K. and Okada, N. (1994) Generality of the tRNA origin of short interspersed repetitive elements (SINEs). Characterization of three different tRNA-derived retroposons in the octopus. *J. Mol. Biol.*, **243**, 25–37.
 68. Tsuchimoto, S., Hirao, Y., Ohtsubo, E. and Ohtsubo, H. (2008) New SINE families from rice, OsSN, with poly(A) at the 3' ends. *Genes Genet. Syst.*, **83**, 227–236.
 69. Gogolevsky, K.P., Vassetzky, N.S. and Kramerov, D.A. (2008) Bov-B-mobilized SINEs in vertebrate genomes. *Gene*, **407**, 75–85.
 70. Singer, D.S., Parent, L.J. and Ehrlich, R. (1987) Identification and DNA sequence of an interspersed repetitive DNA element in the genome of the miniature swine. *Nucleic Acids Res.*, **15**, 2780.
 71. Mochizuki, K., Umeda, M., Ohtsubo, H. and Ohtsubo, E. (1992) Characterization of a plant SINE, p-SINE1, in rice genomes. *Jpn. J. Genet.*, **67**, 155–166.
 72. Motohashi, R., Mochizuki, K., Ohtsubo, H. and Ohtsubo, E. (1997) Structures and distribution of p-SINE1 members in rice genomes. *Theoretical Appl. Genet.*, **95**, 359–368.
 73. Sunter, J.D., Patel, S.P., Skilton, R.A., Githaka, N., Knowles, D.P., Scoles, G.A., Nene, V., de Villiers, E. and Bishop, R.P. (2008) A novel SINE family occurs frequently in both genomic DNA and transcribed sequences in ixodid ticks of the arthropod sub-phylum Chelicerata. *Gene*, **415**, 13–22.
 74. Piskurek, O., Austin, C.C. and Okada, N. (2006) Sauria SINEs: Novel short interspersed retroposable elements that are widespread in reptile genomes. *J. Mol. Evol.*, **62**, 630–644.
 75. Kosushkin, S.A., Borodulina, O.R., Grechko, V.V. and Kramerov, D.A. (2006) A new family of interspersed repeats from squamate reptiles. *Mol. Biol. (Mosk)*, **40**, 378–382.
 76. Deragon, J.M., Landry, B.S., Pelissier, T., Tutois, S., Tourmente, S. and Picard, G. (1994) An analysis of retroposition in plants based on a family of SINEs from *Brassica napus*. *J. Mol. Evol.*, **39**, 378–386.
 77. Deragon, J.M. and Zhang, X. (2006) Short interspersed elements (SINEs) in plants: origin, classification, and use as phylogenetic markers. *Syst. Biol.*, **55**, 949–956.
 78. Lenoir, A., Lavie, L., Prieto, J.L., Goubely, C., Cote, J.C., Pelissier, T. and Deragon, J.M. (2001) The evolutionary origin and genomic organization of SINEs in *Arabidopsis thaliana*. *Mol. Biol. Evol.*, **18**, 2315–2322.
 79. Myouga, F., Tsuchimoto, S., Noma, K., Ohtsubo, H. and Ohtsubo, E. (2001) Identification and structural analysis of SINE elements in the *Arabidopsis thaliana* genome. *Genes Genet. Syst.*, **76**, 169–179.
 80. Daniels, G.R. and Deininger, P.L. (1983) A second major class of Alu family repeated DNA sequences in a primate genome. *Nucleic Acids Res.*, **11**, 7595–7610.
 81. Roos, C., Schmitz, J. and Zischler, H. (2004) Primate jumping genes elucidate strepsirrhine phylogeny. *Proc. Natl Acad. Sci. USA*, **101**, 10650–10654.
 82. Daniels, G.R. and Deininger, P.L. (1991) Characterization of a third major SINE family of repetitive sequences in the galago genome. *Nucleic Acids Res.*, **19**, 1649–1656.
 83. Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M., Wides, R. et al. (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, **298**, 129–149.
 84. Kapitonov, V.V. and Jurka, J. (2003) A novel class of SINE elements derived from 5S rRNA. *Mol. Biol. Evol.*, **20**, 694–702.
 85. Ohshima, K., Koishi, R., Matsuo, M. and Okada, N. (1993) Several short interspersed repetitive elements (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*, **90**, 6260–6264.
 86. Matveev, V., Nishihara, H. and Okada, N. (2007) Novel SINE families from salmonids validate Parahucho (Salmonidae) as a distinct genus and give evidence that SINEs can incorporate LINE-related 3'-tails of other SINEs. *Mol. Biol. Evol.*, **24**, 1656–1666.
 87. Spotila, L.D., Hirai, H., Rekosh, D.M. and Lo Verde, P.T. (1989) A retroposon-like short repetitive DNA element in the genome of the human blood fluke, *Schistosoma mansoni*. *Chromosoma*, **97**, 421–428.
 88. Nisson, P.E., Hickey, R.J., Boshar, M.F. and Crain, W.R. Jr (1988) Identification of a repeated sequence in the genome of the sea urchin which is transcribed by RNA polymerase III and contains the features of a retroposon. *Nucleic Acids Res.*, **16**, 1431–1452.
 89. Carpenter, C.D., Bruskin, A.M., Spain, L.M., Eldon, E.D. and Klein, W.H. (1982) The 3' untranslated regions of two related mRNAs contain an element highly repeated in the sea urchin genome. *Nucleic Acids Res.*, **10**, 7829–7842.
 90. Grechko, V.V., Kosushkin, S.A., Borodulina, O.R., Butaeva, F.G. and Darevsky, I.S. (2011) Short interspersed elements (SINEs) of squamate reptiles (Squam1 and Squam2): structure and phylogenetic significance. *J. Exp. Zool.*, **316B**, 212–226.
 91. Luchetti, A. and Mantovani, B. (2009) Talua SINE biology in the genome of the Reticulitermes subterranean termites (Isoptera, Rhinotermitidae). *J. Mol. Evol.*, **69**, 589–600.
 92. Luchetti, A. and Mantovani, B. (2011) Molecular characterization, genomic distribution and evolutionary dynamics of Short Interspersed Elements in the termite genome. *Mol. Genet. Genomics*, **285**, 175–184.
 93. Smit, A.F. and Riggs, A.D. (1995) MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Res.*, **23**, 98–102.
 94. Yoshioka, Y., Matsumoto, S., Kojima, S., Ohshima, K., Okada, N. and Machida, Y. (1993) Molecular characterization of a short interspersed repetitive element from tobacco that exhibits sequence homology to specific tRNAs. *Proc. Natl Acad. Sci. USA*, **90**, 6562–6566.
 95. Nishihara, H., Terai, Y. and Okada, N. (2002) Characterization of Novel Alu- and tRNA-Related SINEs from the Tree Shrew and Evolutionary Implications of Their Origins. *Mol. Biol. Evol.*, **19**, 1964–1972.
 96. Feschotte, C., Fourrier, N., Desmons, I. and Mouches, C. (2001) 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.*, **18**, 74–84.
 97. Kajikawa, M. and Okada, N. (2002) LINEs mobilize SINEs in the eel through a shared 3' sequence. *Cell*, **111**, 433–444.
 98. Borodulina, O.R. and Kramerov, D.A. (1999) Wide distribution of short interspersed elements among eukaryotic genomes. *FEBS Lett.*, **457**, 409–413.
 99. Lin, Z., Nomura, O., Hayashi, T., Wada, Y. and Yasue, H. (2001) Characterization of a SINE species from vicuna and its distribution in animal species including the family Camelidae. *Mamm. Genome*, **12**, 305–308.