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, polyclonally 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 transposons 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 promotor, 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: rRNA, 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),
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
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
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 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 (SINESs 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. Youri 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-al]. Funding for open access charge: Personal funds.

Conflict of interest statement. None declared.

REFERENCES


