Bioinformatics, 33(5), 2017, 743–745
doi: 10.1093/bioinformatics/btw718
Advance Access Publication Date: 13 December 2016
Applications Note

OXFORD

Genome analysis

SINE_scan: an efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets

Hongliang Mao[1] and Hao Wang[1,2,*]

[1]T-Life Research Center, Department of Physics, Fudan University, Shanghai 200433, People's Republic of China and [2]Department of Genetics, University of Georgia, Athens, GA 30602, USA

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Received on April 11, 2016; revised on October 11, 2016; editorial decision November 8, 2016; accepted on November 9, 2017

Abstract

Motivation: Short Interspersed Nuclear Elements (SINEs) are transposable elements (TEs) that amplify through a copy-and-paste mode via RNA intermediates. The computational identification of new SINEs are challenging because of their weak structural signals and rapid diversification in sequences.
Results: Here we report SINE_Scan, a highly efficient program to predict SINE elements in genomic DNA sequences. SINE_Scan integrates hallmark of SINE transposition, copy number and structural signals to identify a SINE element. SINE_Scan outperforms the previously published *de novo* SINE discovery program. It shows high sensitivity and specificity in 19 plant and animal genome assemblies, of which sizes vary from 120 Mb to 3.5 Gb. It identifies numerous new families and substantially increases the estimation of the abundance of SINEs in these genomes.
Availability and Implementation: The code of SINE_Scan is freely available at http://github.com/maohlzj/SINE_Scan, implemented in PERL and supported on Linux.
Contact: wangh8@fudan.edu.cn
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Short Interspersed Nuclear Elements (SINEs) are transposable elements (TEs) that amplify through a copy-and-paste mode via RNA intermediates. Despite discovered nearly 40 years ago and extensive study in model Eukaryotic organisms (Schmid and Deininger, 1975; Vassetzky and Kramerov, 2013), the computational identification of new SINEs are still challenging.

To date, the only available *de novo* tool specifically designed for SINE identification is a structural based program SINE-Finder (Wenke *et al.*, 2011) that searches the structural signals of plant SINEs and outputs all genomic regions that match these signals. Like other structural based TE discovery tools, SINE-Finder outputs numerous false positive elements even in small genomes. For example, scanning the 116 Mb *Arabidopsis thaliana* and the 200 Mb *Brachypodium distachyon* genomes generated over 500 and 800 putative SINE instances, while manual inspection showed that only 2 (for *A. thaliana*) and 57 (for *B. distachyon*) of these sequences have multiple interspersed full-length copies and thus are likely to be real SINEs (Supplementary Table S1 in Supplemental Information S1). In other words, false discovery rates of SINE-Finder are 99.6% and 93 % in the two genomes. The situation gets worse for large genomes. In the 2Gb B73 maize, SINE-Finder outputs over 3200 putative SINE instances, while only 42 have multiple interspersed full-length copies (false discovery rate = 99%; Supplementary Table S1). Such high false positive rate makes this tool practically incompetent to annotate SINEs in new genomes. It is clear that the information of structural signals is not enough to solve the question of SINE prediction.

The transposition of SINE (and other types of TEs) creates interspersed copies in the genome. Aligning these copies show special

sequence pattern around the insertion site (called transposition hallmark hereafter): the regions belonging to SINE are highly similar while the flanking regions are usually unrelated sequences, except in cases that some SINE copies are located inside duplicated genomic regions derived either by biological reasons (sequence duplication) or artificial results (redundant sequences in unfinished genome draft). Here we argue that transposition hallmark provides powerful information complementary to structural signals and can be used in SINE discovery. Therefore we have developed SINE_Scan, a program that can accurately identify SINE elements in large genomic dataset using transposition hallmark.

2 Methods

SINE_Scan program is composed of three core modules (Supplementary Figs. S1 and S2 in Supplemental Information S1): (i) Candidates collection by *de novo* SINE discovery; (ii) Candidates validation by copy number and transposition hallmark and (iii) Classification and genome-wide annotation. Current SINE_Scan program uses an enhanced version of SINE-Finder as the default tool for candidate collection. The enhanced SINE-Finder can identify all three types (tRNA, 7SLRNA and 5SRNA) of SINEs, while the original SINE-Finder only identifies tRNA type SINE. Reading in genomic sequences, SINE_Scan automatically outputs three files to report SINE landscape in these sequences: a FASTA format file containing representative sequences of SINE families, a FASTA format file containing all reliable SINE copies of these families, and a GFF format file recording positions of these copies. Details of the computational structure of SINE_Scan are described in Section S1 of Supplemental Information S1.

3 Results

3.1 Sensitivity and specificity

Since no benchmarking datasets for plant SINEs have been available, we constructed a dataset (see Section S2) composed of known SINEs and randomly selected non-SINE short TEs from known TEs of rice, sorghum, maize, zebrafish, mouse and human. We obtained 4, 5, 5, 7, 5, 3 known SINE families, respectively. We then collected their non-SINE repeats that are shorter than 1000 bp and randomly chose 20 entries (if possible) from each species. We finally obtained a set of 124 known non-SINE short TEs (Supplementary Table S2). We input the 153 sequences (29 SINEs and 124 non-SINE TEs) into SINE_Scan and found that 122 of the 124 non-SINE TEs and 26 of the 29 SINEs could be correctly identified, which suggested a false positive/negative rate of 1.6% (2/124)/10.3% (3/29), or specificity and sensitivity of 98.4% (122/124) and 89.7% (26/29), respectively. Detail results of this test can be seen in Supplementary Table S2.

SINE_Scan efficiently excluded false sequences in structural search based programs. Our investigation found two major types of false sequences in the major component of the output of SINE_Finder. One type of sequences had low copy number. The other type of sequences, which was predominant in number, had multiple copies in the genome, but they were actually subregions of larger repeats, as evidenced by that the flanking sequences of these SINE candidates are also highly similar (see an example in Supplementary Fig. S3). We compared the prediction results of SINE_Scan and SINE-Finder by running the two programs on 16 plant genome assemblies. The 16 species were rice, sorghum, maize and 13 species that were analyzed by (Wenke *et al.*, 2011). We performed manual verification of SINEs in these genomes and comparison of verified elements to the output of SINE_Scan and SINE-Finder. The results showed that the mean and median false discovery rates of SINE-Finder were 90% and 96%, while the two rates decreased to 4% and 0% in the output of SINE_Scan (Supplementary Table S1).

3.2 SINE_scan discovered numerous new plant SINEs

The application of SINE_Scan to 16 plant and 3 animal genomes (see Supplemental Methods in Supplemental Information S1) resulted in a major reassessment of SINE abundance in many of these genomes (Supplementary Tables S3 and S4). SINE_Scan discovered a number of new SINE families even in organisms in which TEs have been extensively studied. For example, SINE_Scan discovered 4, 21 and 10 new SINE families in rice, barrel medic and poplar. SINE_Scan greatly revised the estimation of SINE abundance in some genomes. For example, SINEs were previously estimated making up 0.04% and 0.02% in *Arabidopsis lyrata* and *Brachypodium distachyon*, respectively. SINE_Scan found 19 and 5 new SINE families, which resulted in reassessment of the abundance of SINEs in the two genomes as 0.23% (almost 6 times than previous estimates) and 0.11% (5 times more than previous estimates). In soybean and cassava, SINE_Scan increased the family number by 19 and 4 folds, and genomic abundance by about 3 and 2 folds, respectively. In tomato and tobacco, two genomes that previous studies did not find SINEs, SINE_Scan first identified 10 and 32 new families, accounting for 0.48% and 0.19% of the two genomes, respectively. Detailed information of families detected by SINE_Scan is presented in Supplementary Table S4. The sequences and genomic distribution can be found in Supplemental Dataset 1.

3.3 Stability and flexibility

We tested the stability of SINE_Scan by running the program under a range of values for 6 important parameters in several species and found that the outputs of SINE_Scan were quite insensitive to the change of values of parameters tested (see details in Supplementary Section S1 and S3; Supplementary Table S6).

SINE_Scan is flexibly designed to meet diverse purposes of SINE annotation and/or validation. The three modules can be used independently or in combination. For examples, when the purpose is to annotate SINEs in a newly sequenced genome, the user is recommended to run the program in the fully automatic mode walking through the 3 modules. Besides using equipped structural search program, users can generate SINE candidates from any other TE identification program (e.g. RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html)) and run the second and third modules to verify and annotate these candidates. When SINE sequences are available and the purpose is to classify and/or find out their representation in the genomic sequences, users can run the third module independently.

In summary, we have shown that SINE_Scan is efficient to identify SINEs in large-scale genomics datasets. It uncovers numerous new SINE families in sequenced plants. It outperforms previous tools by integrating multiple dimensions of information of structural signal, copy-number and hallmark of interspersed insertion to obtain highly reliable SINEs. Our experiments have showed that transposition hallmark is powerful to exclude false positives and discover new elements, thus substantially improves the results of structural-based SINE prediction.

Funding

This work was supported by National Basic Research Program of China (973 Project Grant No. 2013CB34100).

Conflict of Interest: none declared.

References

Schmid,C.W. and Deininger,P.L. (1975) Sequence organization of the human genome. *Cell*, **6**, 345–358.

Vassetzky,N.S. and Kramerov,D.A. (2013) SINEBase: a database and tool for SINE analysis. *Nucleic Acids Res.*, **41**, D83–D89.

Wenke,T. *et al.* (2011) Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *Plant Cell*, **23**, 3117–3128.