SINE_Fisher : A efficient tool for SINEs Identification

Hongliang Mao, Hao Wang

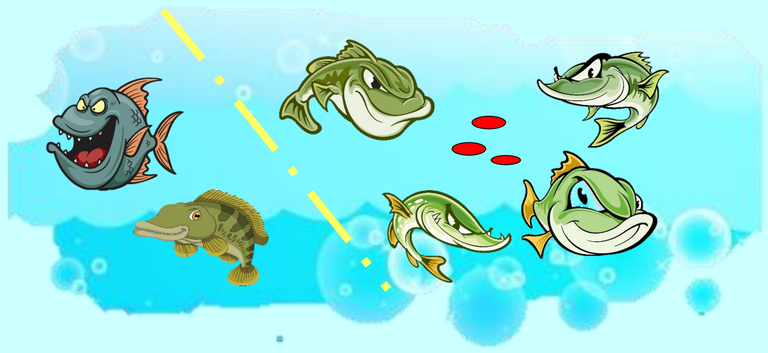
Tlife, Physics Department, Fudan University

A interesting story in genomic sea

1. Hunting DNA fishes in genomic sea

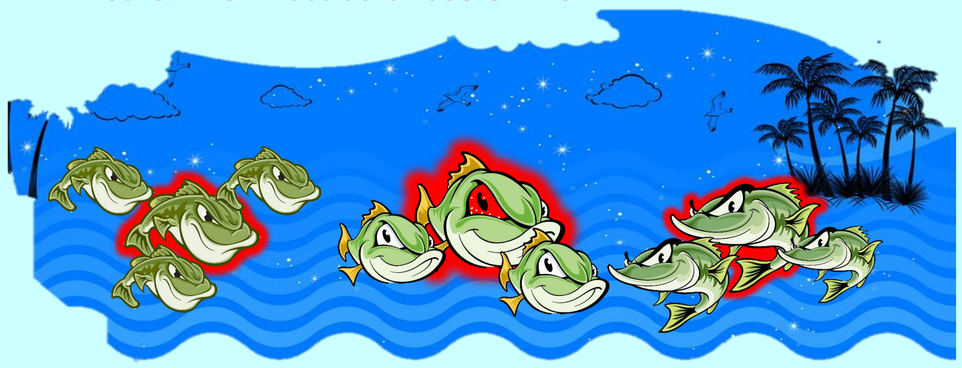
2. Culturing and screening real SINE Fishes

Special fish bait for SINEs



3. Mark SINEs and rehunting in genomic sea

Red SINEs infect colorless SINEs

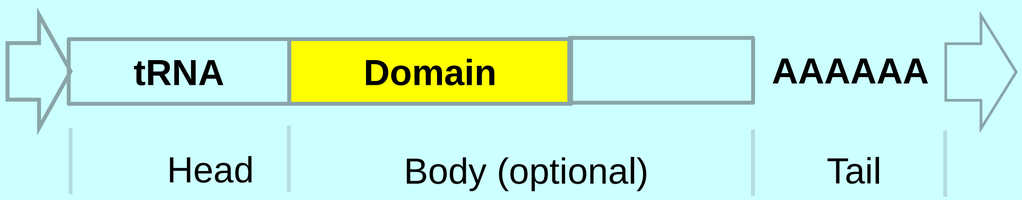


4. Have a rich SINEs' dinner

What are SINEs ?

Since their discovery about 40 years ago, Short Interspersed Nuclear Elements (SINEs), the views upon SINEs have been shifted from selfish DNA to important genomic elements effecting genome organization and function. For example, human *Alu* elements (the first discovered and most famous SINE family) [1], have been found play an important role in disease caused through insertional mutagenesis and non-allelic homologous recombination.

Figure 1 a SINE diagram



Annotation of SINEs

Identification of SINE is challenging. To date, no efficient tools for large-scale genomic data have been developed yet. The only tool available is SINE_Finder [2], which performs pattern search for structural signals of SINE (the existence of A-box, B-box and Target Site Duplicates within a specific stretch of DNA, Figure 1). However, because each and the combination of these structural signals have quite low signal-noise ratio, this structure-based method have very high false positive rate (93% in the case of maize genome (Table 1)).

Table 1 SINE family in the maize genome

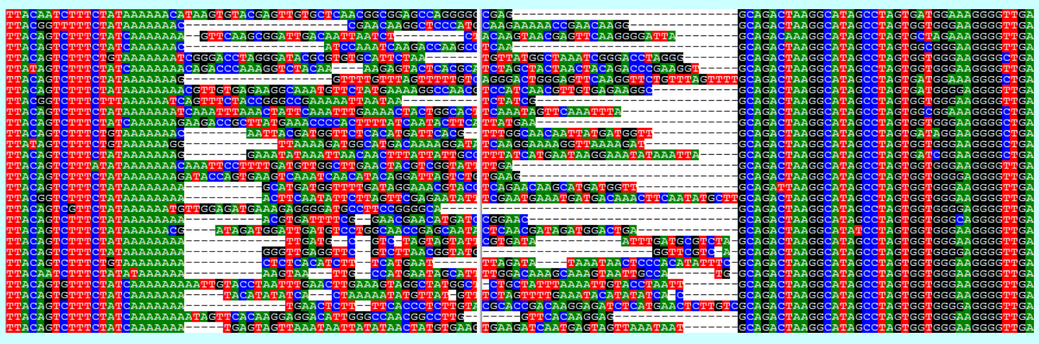
	Number of SINE families	
	SINE-Finder	SINE-Fisher
Total known	6	6
Total identified	1807	122
Verified manually	122	122
New family	116	116
FP	93%	0
FN	0	0

A Brief Introduction to SINE_Fisher

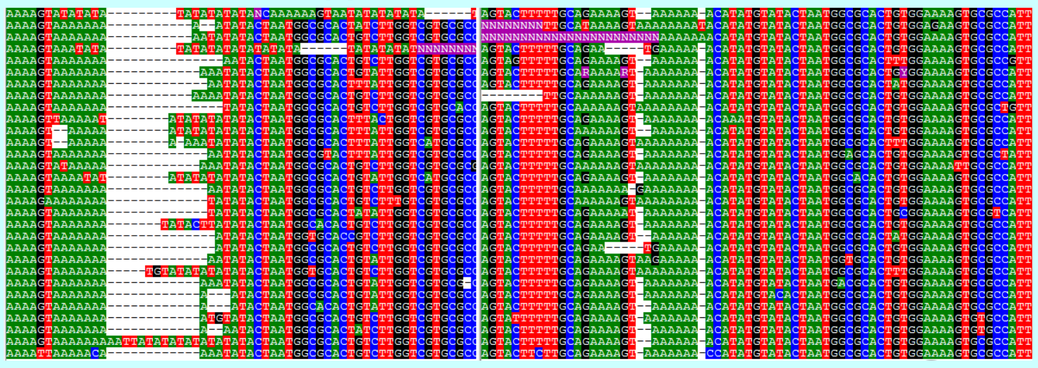
Here we develop an efficient algorithm to identify SINEs, which considers the information of interspersed insertion of SINE in addition to structural signals. When a SINE transposes in the genome, it (1) has multiple copies and (2) insertion sites showed discontinuous quality of alignment because insertion happened at different loci (Figure 2). One of the highlights of this algorithm lies in that it can recognize the insertion of SINEs accurately. The application of our algorithm to maize discovered 116 new SINE families, while previous only 5 families were published (Table 1). More important, this program has both low FP and FN. Take again maize genome as example, all new families identified match gold-standards of SINE (FP=0) and all previously known SINEs can be identified (FN=0).

Figure 2 a good SINE and a bad SINE

(1). a good SINE candidate



(2). a bad SINE candidate



Reference

[1] Schmid CW, Deininger PL (1975). Cell 6: 345–358
[2] Wenke T et. al. (2011). Plant Cell. 23:3117-28