

1 **LTR\_FINDER\_parallel: parallelization of LTR\_FINDER enabling rapid identification of long**  
2 **terminal repeat retrotransposons**

3 Shujun Ou, Ning Jiang\*

4 Department of Horticulture, Michigan State University, East Lansing, MI 48824, USA

5 ORCIDs: 0000-0001-5938-7180 (S.O.); 0000-0002-2776-6669 (N.J.)

6 \*To whom correspondence should be addressed.

7

8 **Abstract**

9 **Summary:** Annotation of plant genomes is still a challenging task due to the abundance of  
10 repetitive sequences, especially long terminal repeat (LTR) retrotransposons. LTR\_FINDER is a  
11 widely used program for identification of LTR retrotransposons but its application on large  
12 genomes is hindered by its single threaded processes. Here we report an accessory program  
13 that allows parallel operation of LTR\_FINDER, resulting up to 8,500X faster identification of LTR  
14 elements. It takes only 72 minutes to process the 14.5 Gb bread wheat (*Triticum aestivum*)  
15 genome in comparison to 1.16 years required by the original sequential version.

16 **Availability:** LTR\_FINDER\_parallel is freely available at

17 [https://github.com/oushujun/LTR\\_FINDER\\_parallel](https://github.com/oushujun/LTR_FINDER_parallel).

18 **Contact:** [jiangn@msu.edu](mailto:jiangn@msu.edu)

19

20 1. Introduction

21 Transposable elements (TEs) are the most prevalent components in eukaryotic  
22 genomes. Among different TE classes, long terminal repeat (LTR) retrotransposons, including  
23 endogenous retroviruses (ERVs), is one of the most repetitive TEs due to their high copy  
24 numbers and large element sizes (Ou and Jiang, 2018). LTR retrotransposons are found in  
25 almost all eukaryotes including plants, fungi, and animals, but are most abundant in plant  
26 genomes (Bennetzen and Wang, 2014). For example, LTR retrotransposons contribute more

27 than 65% and 70% to the genomes of bread wheat (*Triticum aestivum*) and maize (*Zea mays*),  
28 respectively (Ou and Jiang, 2018).

29 Annotation of LTR retrotransposons relies primarily on *de novo* approaches due to their  
30 highly diverse terminal repeats. For this purpose, many computational programs have been  
31 developed in the past two decades. LTR\_FINDER is one of the most popular LTR search  
32 engines and the prediction quality out-performs counterpart programs (Ou and Jiang, 2018).  
33 However, LTR\_FINDER runs on a single thread and is prohibitively slow for large genomes with  
34 long contigs, preventing its application in those species. In this study, we applied the “divide and  
35 conquer” approach to simplify and parallel the annotation task for the original LTR\_FINDER and  
36 observed an up to 8,500 times speedup for analysis of known genomes.

37 2. Methods

38 We hypothesized that complete sequences of highly complex genomes may contain a  
39 large number of complicated nested structures that exponentially increase the search space. To  
40 break down these complicated sequence structures, we split chromosomal sequences into  
41 relatively short segments (1 Mb) and executes LTR\_FINDER in parallel. We expect the time  
42 complexity of LTR\_FINDER\_parallel is  $O(n)$ . For highly complicated regions (i.e., centromeres),  
43 one segment could take a rather long time (i.e., hours). To avoid extended operation time in  
44 such regions, we used a timeout scheme (300 seconds) to control for the longest time a child  
45 process can run. If timeout, the 1 Mb segment is further split into 50 Kb segments to salvage  
46 LTR candidates. After processing all segments, the regional coordinates of LTR candidates is  
47 converted back to the genome-level coordinates for the convenience of downstream analyses.

48 LTR\_FINDER\_parallel is a Perl program that is ready on the go and does not require  
49 any form of installation. We used the original LTR\_FINDER as the search engine which is binary  
50 and also installation free. Based on our previous study (Ou and Jiang, 2018), we applied the  
51 optimized parameter for LTR\_FINDER (-w 2 -C -D 15000 -d 1000 -L 7000 -l 100 -p 20 -M 0.85),  
52 which identifies long terminal repeats ranging from 100 - 7,000 bp with identity  $\geq$  85% and

53 interval regions from 1 - 15 Kb. The output of LTR\_FINDER\_parallel is convertible to the  
54 popular LTRharvest (Ellinghaus, et al., 2008) format, which is compatible to the high-accuracy  
55 post-processing filter LTR\_retriever (Ou and Jiang, 2018).

56 3. Results

57 To benchmark the performance of LTR\_FINDER\_parallel, we selected four plant  
58 genomes with sizes varying from 120 Mb to 14.5 Gb, which are *Arabidopsis thaliana* (version  
59 TAIR10) (Arabidopsis Genome Initiative, 2000), *Oryza sativa* (rice, version MSU7) (International  
60 Rice Genome Sequencing, 2005; Kawahara, et al., 2013), *Zea mays* (maize, version AGPv4)  
61 (Jiao, et al., 2017), and *Triticum aestivum* (wheat, version CS1.0) (International Wheat Genome  
62 Sequencing, et al., 2018), respectively. Each of the genomes was analyzed both sequentially (1  
63 thread) and in parallel (36 threads) with wall clock time and maximum memory recorded.

64

65 Table 1. Benchmarking the performance of LTR\_FINDER\_parallel.

Genome	Arabidopsis	Rice	Maize	Wheat
Version	TAIR10	MSU7	AGPv4	CS1.0
Size	119.7 Mb	374.5 Mb	2134.4 Mb	14547.3 Mb
Original memory (1 thread*)	0.37 Gbyte	0.55 Gbyte	5.00 Gbyte	11.88 Gbyte**
Parallel memory (36 threads*)	0.10 Gbyte	0.12 Gbyte	0.82 Gbyte	17.67 Gbyte
Original time (1 thread)	0.58 h	2.1 h	448.5 h	10169.3 h**
Parallel time (36 threads)	6.4 min	2.6 min	10.3 min	71.8 min
Speed up	5.4 x	48.5 x	2,613 x	8,498 x
# of LTR candidates (1 thread)	226	2,851	60,165	231,043
# of LTR candidates (36 threads)	226	2,834	59,658	237,352
% difference in candidate #	0.00%	0.60%	0.84%	-2.73%

66 \* Intel(R) Xeon(R) CPU E5-2660 v4 @ 2.00GHz

67 \*\* LTR\_FINDER was run on each chromosome; the maximum memory and the total time are  
68 shown.

69

70         Using our method, we observe 5X - 8,500X increase in speed for plant genomes with  
71         varying sizes (Table 1). For the 14.5 Gb bread wheat genome, the original LTR\_FINDER took  
72         10,169 hours, or 1.16 years, to complete, while the multithreading version completed in 72  
73         minutes on a modern server with 36 threads, demonstrating an 8,500X increase in speed (Table  
74         1). Even we analyzed each wheat chromosome separately, the original LTR\_FINDER still take  
75         20 days in average to complete. Among the genomes we tested, the parallel version of  
76         LTR\_FINDER produced slightly different numbers of LTR candidates when compared to those  
77         generated using the original version (0% - 2.73%; Table 1), which is likely due to the use of the  
78         dynamic task control approach for processing of heavily nested regions. Given the substantial  
79         speed improvement (Table 1), we consider the parallel version to be a promising solution for  
80         large genomes.

81

## 82         **Funding**

83         This study was supported by National Science Foundation (IOS-1740874 to N.J.); United States  
84         Department of Agriculture National Institute of Food and Agriculture and AgBioResearch at  
85         Michigan State University (Hatch grant MICL02408 to N.J.).

86         *Conflict of Interest: none declared.*

87

## 88         **References**

89         Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant  
90         *Arabidopsis thaliana*. *Nature* 2000;408(6814):796-815.

91         Bennetzen, J.L. and Wang, H. The Contributions of Transposable Elements to the Structure,  
92         Function, and Evolution of Plant Genomes. *Annu. Rev. Plant Biol.* 2014;65(1):505-530.

93         Ellinghaus, D., Kurtz, S. and Willhoeft, U. LTRharvest, an efficient and flexible software for *de*  
94         *novo* detection of LTR retrotransposons. *BMC Bioinformatics* 2008;9(1):18.

95 International Rice Genome Sequencing, P. The map-based sequence of the rice genome.

96 *Nature* 2005;436(7052):793-800.

97 International Wheat Genome Sequencing, C., *et al.* Shifting the limits in wheat research and

98 breeding using a fully annotated reference genome. *Science* 2018;361(6403).

99 Jiao, Y., *et al.* Improved maize reference genome with single-molecule technologies. *Nature*

100 2017;546:524-527.

101 Kawahara, Y., *et al.* Improvement of the *Oryza sativa* Nipponbare reference genome using next

102 generation sequence and optical map data. *Rice* 2013;6(1):1-10.

103 Ou, S. and Jiang, N. LTR\_retriever: A Highly Accurate and Sensitive Program for Identification

104 of Long Terminal Repeat Retrotransposons. *Plant Physiol.* 2018;176(2):1410-1422.

105