LTR Annotator: Automated Identification and Annotation of LTR Retrotransposons in Plant Genomes

Frank M. You[1, 2*], Sylvie Cloutier[2, 3], Yunfeng Shan[1], Raja Ragupathy[2]

[1] Cereal Research Centre, Agriculture and Agri-Food Canada, 101 Route 100, Unit 100 Morden, MB, R6M 1Y5, Canada.
[2] Department of Plant Science, University of Manitoba, 66 Dafoe Road, Winnipeg, MB, R3T 2N2, Canada.
[3] Eastern Cereal and Oilseed Research Centre, Agriculture and Agri-Food Canada, 960 Carling Avenue, Ottawa, ON K1A 0C6, Canada.

* Corresponding author. Tel.: +1-204-822-7525; email: Frank.You@agr.gc.ca
Manuscript submitted January 6, 2015; accepted March 8, 2015.
doi: 10.17706/ijbbb.2015.5.3.165-174

Abstract: Long Terminal Repeat transposable element (LTR) is a major type of mobile elements ubiquitous in eukaryotic genomes. They account for a major proportion of many plant genomes and have a prominent impact on the evolution of genome size, structure and function. Although some bioinformatics tools for *de novo* LTR identification from genome sequences have been developed, an automated and standardized software tool for both LTR identification and annotation would be valuable and essential for comparative analysis of sequenced plant genomes. We present here a Java-based pipeline tool, called LTR Annotator, for automatically and consistently performing genome-wide *de novo* identification and annotation of LTRs of plant genome sequences. The pipeline first identifies LTRs using both LTR_FINDER and LTR harvest, then performs intensive annotations, and finally sweeps out potentially false-positive LTRs. The pipeline was evaluated using the well curated *Arabidopsis* genome. High sensitivity (>0.9) was obtained by using LTR harvest or LTR harvest+LTR_FINDER. Ten potentially new intact LTRs were detected. This pipeline provides a comprehensive tool to perform comparative analysis of LTRs for plant genomes, delivering annotated genomic resources for epigenetic and other studies. LTR Annotator is free and available upon request.

Key words: Retrotransposon, LTR, plant genome, LTR annotator, annotation.

1. Introduction

Transposable elements (TEs) are DNA fragments that are capable of moving from one location to another in a genome. They are classified based on transposition mechanisms into two categories: retrotransposons (Class I) and DNA transposons (Class II) [1]. While TEs account for ~14 to ~84% of entire plant genomes, Class I elements account for more than 80% of the TEs in half of the sequenced plant genomes and as much as 96% in banana and 98% in tomato [2]. Among Class I elements, Long Terminal Repeat transposon (LTR), ubiquitous in eukaryotes, is a major component, accounting for an average of 82% of Class I elements.

Because of their significant structural features, identification of LTR retroelements can be carried out using a *de novo* approach. Several programs have been released for identifying full-length or intact LTRs, such as LTR_STRUCT [3], LTR_PAR [4], FIND_LTR [5], LTR_FINDER [6] and LTR harvest [7]. These tools take into account several major characteristics of LTRs such as the size range of intact LTRs, the distances between two LTRs of intact elements, the presence of target site duplications (TSDs) at each terminal region,

the presence of critical sites for reversing transcribing elements for transposition such as the primer binding site (PBS) and the poly purine tract (PPT), and the identity percentage between two LTRs. In addition, LTR digest [8] (an internal domain annotation tool) and LTR shift [9] (a graphic viewer of LTRs and their families predicted by LTR harvest [7]) were developed. A systematic evaluation of those tools using sequences of the X chromosome of the *Drosophila melanogaster* genome [10] indicated that LTRharvest, FIND_LTR and LTR_FINDER outperformed other tools in detecting LTRs. However, common concerns for these three tools are the high rate of false positives [10], and varying criteria for TE annotation [1]. The lack of recognized guidelines results in a wide variety of software tools generating different results which impede our ability to perform comparative analyses of TE component of multiple genomes based on uniform criteria without performing re-annotation.

Here, we report on a comprehensive LTR annotation pipeline, called LTR Annotator, which combines *de novo* LTR identification and annotation. This pipeline amalgamates *de novo* and homology based approaches together to identify new LTR elements and filter out false positives. We tried to develop a standardized LTR annotation procedure to facilitate comparative analysis of genomes and to provide annotated genomic resources that can be exploited for understanding the epigenetic impact of TEs on the gene space.

2. Design and Implementation of the LTR Annotator Pipeline

The LTR Annotator pipeline was designed to include three major components: LTR identification, LTR annotation and result summarization as illustrated in Fig. 1.

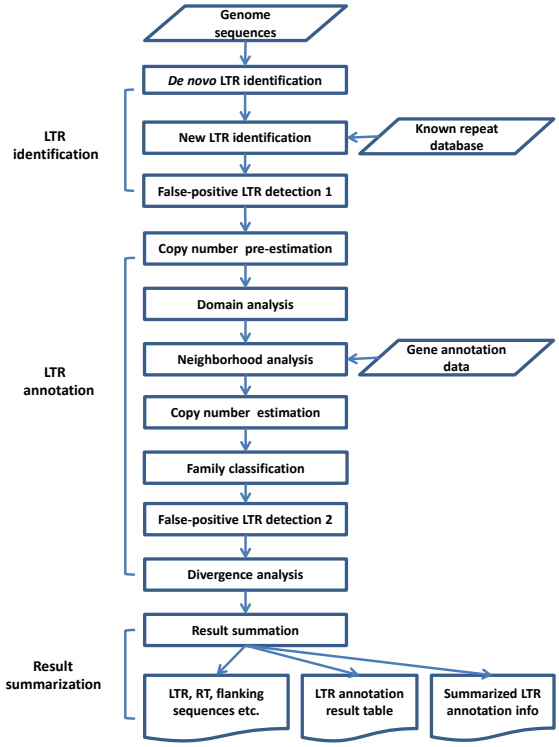


Fig. 1. Workflow of the LTR Annotator pipeline.

2.1. LTR Identification

2.1.1. De novo LTR identification

The previously published third-party programs LTR_FINDER and LTR harvest were integrated into the pipeline for *de novo* identification of candidate LTR elements. The main reason to choose these two

programs was because of the low false-positive rate of LTR_FINDER and the high sensitivity of LTR harvest [10]. A set of similar parameters were applied while running these two programs. While all parameters could be modified by the user in the configuration file, the default parameters for both programs of 100 bp minimum and 3,000 bp maximum LTR length, and 0.85 similarity between two LTRs were retained. Some specific parameters for LTR harvest were used, including "–xdrop 37 –motif tgca –mootifmis 1 –mintsd 2". While launching the pipeline, either or both of the programs can be selected in the parameter settings of the LTR identification modules. When both programs are selected, unique LTR candidates are identified from all LTR candidates based on their predicted coordinates and only one copy is retained.

2.1.2. New LTR identification

Homology search of LTR candidates to a known annotated repeat database is used to identify new LTR candidates and to characterize LTRs. Two repeat databases, MIPS-REdat 9.0 (http://mips.helmholtz-muenchen.de/plant/recat/) and TREP (Triticeae repeats) (http://wheat.pw.usda.gov/ITMI/repeats) were adopted as common reference libraries. MIPS-REdata v9.0 contains more than 42,000 repeat elements integrated from multiple repeat databases, including entries of both Class I retrotransposons and Class II DNA transposons as well as super-family information. TREP is a well annotated repeat database for the Triticeae tribe of the grass family which has family classification information of all Class I and II entries from wheat, barley and rye. The current version (v1.0) of the pipeline can only accept the header line format of these two databases (fasta files).

Initially, LTRs of all intact LTR candidates are searched for their homology against a user-specified repeat database using BLASTn with an E-value threshold of 1e-30. The annotation information of the best hit is assigned to each LTR candidate. The best hit is defined as the top hit generated by the BLAST algorithm with at least 80% alignment identity over 80% of the LTR length or database entry length with a minimum of 80 bp of aligned sequences. This was originally referred to as the 80-80-80 rule [1]. The reference annotation information is considered for LTR super-family assignment and family classification. The LTR candidates without any hits to the reference repeat database are classified as potential "new" LTR elements and are subjected to further scrutiny.

2.1.3. Detection of false-positive LTRs

Many false-positive LTRs are generated from the structure-based *de novo* LTR identification software, such as LTR harvest [10]. One of the critical tasks of the pipeline is to identify and remove false-positive LTRs from the LTR candidates. Duplicated genes and tandemly repeated DNA transposons are major sources of false calls. The detection and removal of these false-positive LTRs are carried out in two separate steps (Fig. 2). The first step entails the removal of tandem DNA transposons misdiagnosed as LTRs. The second confirmation consists of a module incorporated immediately after copy number estimation, internal domain analysis and LTR neighborhood analysis. In this module, both copy number and homology to known genes are taken into consideration. Wicker *et al*. [1] suggested that a potential TE should have at least five copies in a genome. If the copy number of LTR candidates belong to unknown super-families inferred based on homology search against reference repeat databases is less than 5 [1] or 4 [11] or, if two LTR candidates are annotated as members of the same gene family, then those LTRs are considered false-positives. However, the evolution of LTRs is a dynamic process. LTRs evolve from somewhere and their initial copy number will be lower (e.g., 4) before they further multiply. Thus this minimum copy number of a true LTR as an input parameter of the pipeline can be adjusted in the configuration based on the user's considerations. The candidates remaining after removal of the so-identified false-positives, are retained for final summary and the results are exported.

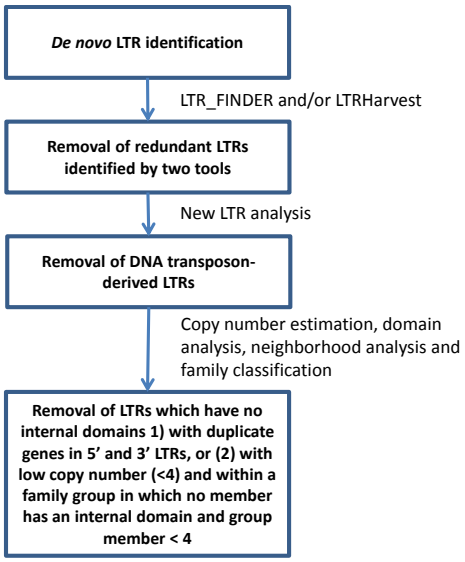


Fig. 2. Identification and removal of putative false-positive LTRs.

2.2. LTR Annotation

The LTR annotation component integrates multiple algorithms and third party tools to obtain accurate annotations. Those integrated include LTR internal domain analysis, copy number estimation, super-family assignment, family classification, LTR-pair based divergence analysis and LTR neighbourhood analysis.

2.2.1. Domain analysis

A typical intact LTR element comprises the *gag* and *pol* internal domains. The *gag* polyprotein is processed into capsid, nucleocapsid, spacer peptide and p6 proteins. Similarly, the *pol* polyprotein includes reverse transcriptase (RT), RNAse H, integrase (INT) and protease. Hidden Markov Model (HMM) profile-based prediction of protein domains in the internal region flanked by LTRs was carried out to identify putative intact elements.

2.2.2. Copy number estimation

Sequence identity (>80%) of LTRs and/or internal regions was used to identify members belonging to the same family, as suggested earlier [1]. BLASTn search of identified LTR elements against whole genome sequences is performed to identify intact, solo or fragmental LTRs [11]. The copy number provides an indication of LTR activity during genome evolution, and also serves to discriminate between true and false LTR-like structures. Some *de novo* identified LTR elements may not be considered if their copy number is less than the cutoff value of 5 [1] or 4 [11]. All three forms, namely intact elements, solo LTRs or fragments of complete elements, are considered for the copy number estimation per family. This is necessitated by the current accepted model of genome evolution with amplification bursts of LTR elements followed by elimination of repeat sequences (Increase/decrease model [12]). Deletion is mediated by illegitimate homologous recombination between similar LTRs of either same or different elements (leading to solo LTRs) or other direct repeats (leading to internal deletions). Percentages of different forms of LTR elements (intact, solo LTR and fragment) in whole genome are also calculated.

2.2.3. Super-family and family classification

There are two strategies implemented to classify elements into super-families and families. Firstly, deduced RT sequences are aligned using MAFFT [13] to infer phylogenetic relationships. Another LTR clustering approach provided with SiLiX [14] dramatically improves clustering performance. Secondly, the

order of *gag* and *pol* domains in intact candidates is taken into consideration for assigning super-families [1].

2.2.4. Divergence analysis

Divergence analysis is performed through assessing the evolutionary divergence of 5' and 3' LTRs of individual elements. Insertion times of LTR copies are estimated using the molecular clock proposed for LTR element evolution [15]. Left and right LTRs of intact elements are aligned using the Needleman–Wunsch algorithm (global alignment) implemented in BioJava 3 [16]. The evolutionary distance between two LTRs is calculated based on nucleotide substitutions corrected with the Kimura II parameter model of nucleotide evolution, which accounts for multiple substitutions per site [17]. The number of substitutions (k) between two LTRs of an element can be converted into insertion time (t) in million years (MY) by $t = k/(2r)/10^{-6}$, where r is the substitution rate. The average pipeline default substitution rate of 1.3×10^{-8} substitutions per synonymous site per year [18] was used, although this value can be user-defined in the configuration file.

2.2.5. Neighbourhood analysis

The purpose of neighbourhood analysis is to provide clues to examine whether LTRs play a role in regulation and expression of a flanking or neighbouring gene(s). Thus positional information of genes around an intact LTR element is collected as outlined in Fig. 3. (A) Extraction of 5kb upstream and 5kb downstream flanking regions of an intact putative LTR. (B) Detection of the presence of genes in the 5kb flanking regions. The distance between a gene and an LTR is calculated on both sides if a gene is detected. (C) Identification of genes either overlapping with or present inside of an LTR. This step helps to remove false positives when duplicated genes overlap with LTRs or the copy number of the putative LTR element is very low. This information linking LTRs to gene annotation will be useful to predict epigenetic impact of LTRs on the gene space.

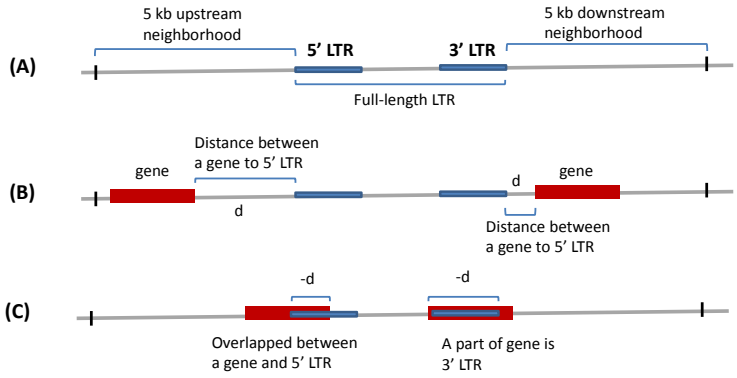


Fig. 3. Diagram of LTR neighbourhood analysis.

2.2.6. S/I ratio estimation

Homologous ectopic recombination between 5' and 3' LTRs is a main mechanism of element elimination, leading to the formation of solo-LTRs [19]. Hence, the ratio of solo-LTRs (S) to intact (complete) elements (I) for a particular family with multiple members would indicate the rate of elimination of amplified copies and hence the dynamics of genome size and structure. An S/I ratio smaller than one suggests a slow rate of removal compared to amplification while a ratio greater than one indicates a high elimination rate. The unimodality test for intact LTR length distribution of each family is performed as previously suggested [11]

using the R module DIPTEST to test for unimodal distribution. If it is unimodal, its mode may correspond to the size of the most frequent paralog found in the genome for the family. One could hypothesize that this mode corresponds to the original master copy. Thus, paralogs with a smaller size than the mode would correspond to deleted copies, whereas paralogs with a larger size should correspond to copies with nested insertions of other TEs [11].

2.3. Result Summarization

After the core modules' run is completed, the result summarization component gives rise to a list of files to be exported, including a log file and fasta files of sequences for known and unknown super-families. An output file with inferred phylogenetic tree can be visualized with any third party software commonly used for this purpose, for instance MEGA6. The XX_LTR_annotation_summary.txt and XX_LTR_annotation_table.txt are two summary files that include summarized information of LTR identification and annotation. All exported files are self-explanatory and their contents are described in the user's guide accompanying the software release.

2.4. Implementation of the Pipeline

The pipeline is a command-line based tool. All data and intermediate/final results are stored in a Java-based relational database (HSQLDB, http://hsqldb.org). The software is written in Java and Perl. To improve pipeline performance for data mining of large genomes, Java multiple threads are implemented to simultaneously perform various analyses, and consequently speeding up data processing in any types of machines with single core or multicore CPUs as well as clusters. The number of threads in the configuration file is user-defined. Some third party software, libraries and databases are incorporated into the pipeline, such as LTR_FINDER, LTR-Harvest, BLASTn, Biojava, Bioperl, MAFFT, tRNA sequences for PBS and PBT identification, HMM-based profiles of protein domains, TREP, Repbase and mipsREdat. A gene annotation (GFF format) file is needed for complete analysis of each genome under study.

3. Results and Discussion

The flowering plant *Arabidopsis thaliana* is an important model system for identifying genes and determining their functions. It was the first completely sequenced plant species in 2000 [20] and it has been well annotated since then. We therefore used its most recent annotation (TAIR10) to evaluate the pipeline. From the TAIR database (https://www.arabidopsis.org/), a total of 5,962 LTR sequences with a size range of 12-31,019 bp (1,382 ± 2,188 bp) were obtained, accounting for 6.9 % of the genome assembly. According to self-BLAST of the LTR sequences, we only found 167 intact LTR sequences, but noticed that many intact LTRs were represented in separate LTR fragments. We re-analyzed all these neighboring LTR fragments and aligned them into intact LTRs. Subsequently 638 intact LTRs were obtained and used for pipeline evaluation.

We integrated two *de novo* LTR identification algorithms (software) into the pipeline. A user has the option to choose a single algorithm or a combination of both. To evaluate the performance of the pipeline, we compared the results obtained using three methods (Table 1). As evaluated in Ref. [4], LTR harvest and LTR-FINDER+LTR harvest generated a large number of false-positive LTRs which required further removal through annotations. Most false positives resulted from duplicated genes and, as such, possessed no significant internal protein domains. Although LTR-FINDER detected fewer false positives, some true positives were swept out or remained undetected in the *de novo* identification. Consequently, LTR_FINDER should not be used alone. In this case, using LTR harvest and the combination of the two methods resulted in sensitivity (0.91) considerably higher than LTR-FINDER alone (0.52).

Table 1. LTR Identification and Annotation Using the LTR Annotator Pipeline for *Arabidopsis thaliana* Genome Assembly (TAIR10)

Item	LTR_FINDER+LTR harvest	LTR-FINDER	LTR harvest
Total LTRs identified initially	2,837	670	2,619
LTRs after filtering	668	444	728
Detected LTRs in TAIR10	588	333	579
Undetected LTRs in TAIR10	50	305	59
Detected LTRs not in TAIR10	150	111	149
Sensitivity	0.92	0.52	0.91

The total number of intact LTRs in the curated data (TAIR10) is 638.

Although the *Arabidopsis* genome is well annotated, we still identified ~150 intact LTRs with significant internal domain structures that were not included in the TAIR10 annotation. Ten of the identified LTRs were not aligned to any LTR sequences in the current LTR database and all detected LTR sequences in *Arabidopsis* at the DNA level with low homology similarity in the RT domain with existing LTRs (Table 2). Six of these LTRs may have diverged and generated new copies recently (< 1 million years ago, MYA). All ten LTRs belonged to the *Copia* super-family. Further phylogenetic analysis of the RT proteins showed that ten LTRs may be grouped into five families with four copies for one family (a) and three copies for another family (b) (Fig. 4, Table 2). In the Fig. 4, Boot-strap values are labelled based on 1000 replicates. Ten intact LTRs are classified into five families (a-e).

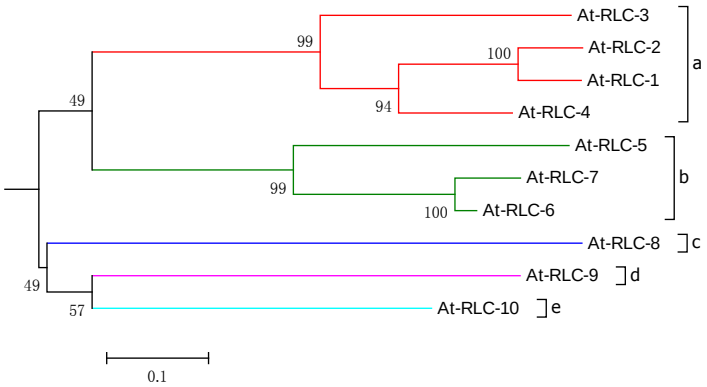


Fig. 4. Neighbour-Joining tree of 10 newly identified LTRs based on RT protein sequences.

Table 2. New LTRs Identified from the *Arabidopsis* Genome

ID	Chr	Coordinates		LTR length (bp)	R[a]	Internal domain structure	F[b]	Diversity (MYA)
		Start	End					
At-RLC-1	1	17023916	17029257	5,342	0.97	GAG-INT-RT	a	1.3
At-RLC-2	3	16957465	16962731	5,267	0.97	GAG-INT-RT	a	1.1
At-RLC-3	4	6588726	6593808	5,083	0.99	GAG-INT-RT	a	0.4
At-RLC-4	4	10992996	10998058	5,063	0.98	INT-RT	a	0.6
At-RLC-5	1	9476002	9480634	4,633	0.98	GAG-INT-RT	b	0.6
At-RLC-6	3	11016111	11020804	4,694	0.94	GAG-INT-RT	b	2.3
At-RLC-7	1	22100644	22105329	4,686	0.9	GAG-INT-RT	b	4.5
At-RLC-8	2	9121958	9125818	3,861	1	GAG-INT-RT	c	0.1
At-RLC-9	3	11122707	11127674	4,968	0.99	GAG-INT-RT	d	0.3
At-RLC-10	4	2188979	2194314	5,336	1	INT-RT	e	0.2

[a] R: Similarity between two LTRs; [b] F: Family. All LTRs are grouped into the *Copia* super-family. Chr: chromosome; GAG: the genomic region encoding the capsid proteins; INT: integrase; RT: reverse transcriptase; MYA: million years ago.

The computing performance of LTRAnnotator depends on the complexity and the size of the genomes processed. The total running time (hours estimated based on a workstation with 32GB, Intel® Core™ i7-2920XM CPU @2.50GHZ) linearly increased with the genome size (Fig. 5). In the multiple steps of the pipeline, estimation of copy numbers, *de novo* detection of LTRs, domain analysis, neighborhood analysis and LTR super-family classification took approximately 57%, 27%, 9%, 2% and 2% of the total running time, respectively. Copy number estimation contributed to the most running time. Because, in this module, all copies of LTRs are compared with the entire genome sequence. We used KLAST (http://www.korilog.com) in this pipeline for sequence comparison which dramatically improved performance by nearly 10-fold over BLAST.

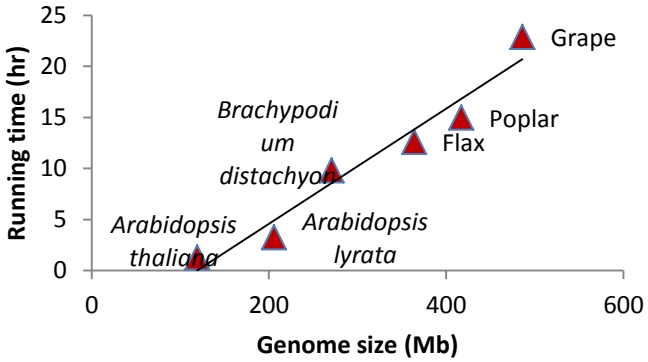


Fig. 5. Computing performance of LTR Annotator in relation to genome size.

4. Conclusions

The LTR Annotator pipeline has been developed to combine *de novo* LTR detection and LTR annotation together with removal of potential false-positive LTRs. This pipeline delivers detailed results of LTR annotations such as internal domains, PBS, PPT sequences, copy numbers of identified intact elements and solo-LTRs, divergence estimates based on LTR pairs and presence of genes in the neighborhoods of identified LTR elements, which allows us to perform comparative analyses of LTRs across multiple sequenced plant genomes using common parameters. This pipeline has been applied to comparative analysis of LTRs in about 40 sequenced plant genomes (http://phytozome.com/). The results will be published separately.

Acknowledgment

This work was supported by the Total Utilization Flax GENomics (TUFGEN) project funded by Genome Canada and other stakeholders and the A-base project J-000066 funded by Agriculture and Agri-Food Canada. We sincerely thank Andrzej Walichnowski for manuscript editing.

References

[1] Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., *et al*. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, *8(12)*, 973-982.

[2] Ragupathy, R., You, F. M., & Cloutier, S. (2013). Arguments for standardizing transposable element annotation in plant genomes. *Trends in Plant Science*, *18(7)*, 367-376.

[3] McCarthy, E. M., & McDonald, J. F. (2003). LTR_STRUC: A novel search and identification program for

LTR retrotransposons. *Bioinformatics*, *19(3)*, 362-367.

[4] Kalyanaraman, A., & Aluru, S. (2006). Efficient algorithms and software for detection of full-length LTR retrotransposons. *Journal of Bioinformatics and Computational Biology*, *4(2)*, 197-216.

[5] Rho, M., Choi, J. H., Kim, S., Lynch, M., & Tang, H. (2007). *De novo* identification of LTR retrotransposons in eukaryotic genomes. *BMC Genomics*, *8*, 90.

[6] Xu, Z., & Wang, H. (2007). LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, *35*(Web Server issue), 265-268.

[7] Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, *9*, 18.

[8] Steinbiss, S., Willhoeft, U., Gremme, G., & Kurtz, S. (2009). Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Research*, *37(21)*, 7002-7013.

[9] Steinbiss, S., Kastens, S., & Kurtz, S. (2012). LTRsift: A graphical user interface for semi-automatic classification and postprocessing of de novo detected LTR retrotransposons. *Mobile DNA*, *3(1)*, 18.

[10] Lerat, E. (2010). Identifying repeats and transposable elements in sequenced genomes: How to find your way through the dense forest of programs. *Heredity (Edinb)*, *104(6)*, 520-533.

[11] El Baidouri, M. & Panaud, O. (2013). Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. *Genome Biology and Evolution*, *5(5)*, 954-965.

[12] Vitte, C., & Panaud, O. (2005). LTR retrotransposons and flowering plant genome size: Emergence of the increase/decrease model. *Cytogenetic and Genome Research*, *110(1-4)*, 91-107.

[13] Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30(4)*, 772-780.

[14] Miele, V., Penel, S., & Duret, L. (2011). Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics*, *12*, 116.

[15] SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y., & Bennetzen, J. L. (1998). The paleontology of intergene retrotransposons of maize. *Nature Genetics*, *20(1)*, 43-45.

[16] Prlic, A., Yates, A., Bliven, S. E., Rose, P. W., Jacobsen, J., Troshin, P. V., *et al*. (2012). BioJava: An open-source framework for bioinformatics in 2012. *Bioinformatics*, *28(20)*, 2693-2695.

[17] Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, *16(2)*, 111-120.

[18] Ma, J., Devos, K. M., & Bennetzen, J. L. (2004). Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Research*, *14(5)*, 860-869.

[19] Vitte, C., Panaud, O., & Quesneville, H. (2007). LTR retrotransposons in rice (Oryza sativa, L.): Recent burst amplifications followed by rapid DNA loss. *BMC Genomics*, *8*, 218.

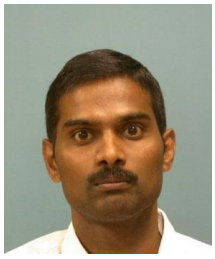
[20] Initiative, A. G. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, *408(6814)*, 796-815.

Frank You is a research scientist in bioinformatics and genomics at Cereal Research Centre, Agriculture and Agri-Food Canada (AAFC), Canada and also an adjunct professor of Department of Plant Science, University of Manitoba. He obtained his MSc and PhD degrees in plant genetics and breeding with statistical genetics in Nanjing Agricultural University, and his bachelor degrees in agronomy in Sichuan Agricultural University and in computer science in University of Manitoba. He has expertise in computational biology and bioinformatics, statistical genetics, and plant genetics and breeding. His research interests include plant comparative and statistical genomics, quantitative genetics, genome

assembly and annotation of complex genomes, gene expression and microarray data analysis, physical mapping and data analysis, high-throughput molecular marker design and development, and bioinformatics software development. He has contributed more than 80 peer-reviewed papers in scientific journals.

Sylvie Cloutier is a research scientist at the Eastern Cereal and Oilseed Research Centre of Agriculture and Agri-Food Canada located in Ottawa, Canada. She is also an adjunct professor at the Universities of Manitoba, Guelph and Ottawa. She obtained her PhD in plant molecular genetics from the Université de Montréal, her MSc in biotechnology from the University of Guelph and her BSc Appl in agronomy from Université Laval. After one year as a visiting fellow at the Cereal Research Centre of AAFC in Winnipeg, she joined AAFC as a scientist, first at CRC in Winnipeg and recently at ECORC in Ottawa. Her research interests are mostly in flax and wheat genomics and epigenetics including structural genomics, quantitative genetics, genome analyses, small RNAs, biotic and abiotic stress resistance. She has published nearly 80 peer-reviewed scientific articles in refereed journals and seven book chapters and review articles. She co-led an international project entitled TUFGEN and received the 2013 Rosemary Davis award for leadership in agriculture.

Yunfeng Shan is a bioinformatics researcher. He recently worked on a database with GMOD Chado and analysis web application with Tripal. He has experiences in De Novo assembly, SNP identification and RNA-seq analysis. He developed the computer program of maximum gene-support tree in C (gsupport.c) for genome-scale phylogenetic analysis. His research interests span genomics, bioinformatics, air pollution impacts and ecophysiology.



Raja Ragupathy is a research scientist at the National Rice Research Institute of the Indian Council of Agricultural Research in Cuttack, India. He obtained his PhD from the University of Manitoba in Canada and his MSc and BSc from Tamil Nadu Agricultural University in India. Upon graduation with his PhD, he spent several years in Manitoba as a visiting fellow in a Canadian government laboratory and as a research associate at the University of Manitoba. He was also an assistant professor of genetics and plant breeding at Pandit Jawaharalal Nehru College of Agriculture and Research Institute in India for more than four years. He has published 13 peer-reviewed scientific articles and one book chapter. He has also received many academic awards including the Manitoba Graduate scholarship of the Government of Manitoba for outstanding international students, the William B. Malchy fellowship and the Clarence-Bougardus Sharp memorial awards.