BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

MITE Digger, an efficient and accurate algorithm for genome wide discovery of miniature inverted repeat transposable elements

Guojun Yang

Abstract

Background: Miniature inverted repeat transposable elements (MITEs) are abundant non-autonomous elements, playing important roles in shaping gene and genome evolution. Their characteristic structural features are suitable for automated identification by computational approaches, however, *de novo* MITE discovery at genomic levels is still resource expensive. Efficient and accurate computational tools are desirable. Existing algorithms process every member of a MITE family, therefore a major portion of the computing task is redundant.

Results: In this study, redundant computing steps were analyzed and a novel algorithm emphasizing on the reduction of such redundant computing was implemented in MITE Digger. It completed processing the whole rice genome sequence database in ~15 hours and produced 332 MITE candidates with low false positive (1.8%) and false negative (0.9%) rates. MITE Digger was also tested for genome wide MITE discovery with four other genomes.

Conclusions: MITE Digger is efficient and accurate for genome wide retrieval of MITEs. Its user friendly interface further facilitates genome wide analyses of MITEs on a routine basis. The MITE Digger program is available at: http://labs.csb.utoronto.ca/yang/MITEDigger.

Background

Miniature inverted repeat transposable elements (MITEs) are short non-autonomous transposable elements (TEs) that move by cut-and-paste mechanisms [1-3]. They do not produce transposases, proteins that mobilize TEs, and are therefore dependent on those produced by autonomous elements for transposition [4,5]. Compared to typical cut-and-paste transposons, MITE families often have high copy numbers, and transposition of these elements generates widespread genomic variations [6-8]. Due to their small sizes, MITE insertions are much less disruptive to genes than the larger elements. Therefore, they can often be found in genic regions, introducing phenotypical changes in some cases. For example, an insertion of a *Stowaway* MITE named *dTstu*1 in the flavonoid 3′,5′-hydroxylase (F3′5′H) gene of a potato leads to red pigmentation, and a *mPing* insertion in the rice *Hd*1 gene results in changes in flowering time [9-11]. Most MITE insertions may not cause phenotypical changes, but rather they alter gene expression levels and epigenetic profiles that may contribute to the overall fitness of the organisms under certain conditions [6,12-16]. While understanding how MITEs transpose to achieve high copy numbers can further our knowledge on their influence on genome evolution and provide MITE based genetic tools [4,17-19], genome wide identification and characterization of MITE families broaden our views on different types of MITEs and the scale of their activity and amplification during evolution [6,20-23]. New MITE families may become better candidates for studies of their transposition and amplification as well as for genetic markers.

MITEs were first discovered from the genetic variation caused by an insertion at the maize *wx*-B2 locus [24]. Computational approaches were employed to assist the characterization of MITEs with the increasingly available genomic sequences in databases around early 1990s [25]. Due to their well defined structural features including small size (50–800 bp), terminal inverted repeats (TIRs) and target site duplications (TSDs), the task to discover

Correspondence: gage.yang@utoronto.ca
Department of Biology, University of Toronto Mississauga, Mississauga ON L5L 1C6, Canada

MITE families at genomic levels is suitable for automation. However, the complexity of higher eukaryotic genomes presents a major challenge for such automation. The TSDs and TIRs of MITEs are very short sequences that can occur at a high frequency by chance, resulting in a large number of false output entries that need to be manually analyzed [26-28]. Automated genome wide identification of MITEs can be time consuming because of the large sizes of higher eukaryotic genomes and high TE contents. Such computing tasks are often demanding on computing resources such as the number of CPUs and the amount of RAM.

The program FINDMITE was developed and used for the discovery of eight novel MITE families in the malaria mosquito *Anopheles gambiae* genome from a sequence database containing short entries [28]. Its input parameters include a predefined sequence or size of TSD, the length of TIRs, and the minimal distance between TIRs. All sequences satisfying the parameters are retrieved and processed. The program MUST, MITE Uncovering SysTem, is based on string matching to identify candidate TIR structures followed by checking the presence of a flanking TSD pair [29]. All candidates are retrieved and grouped. For genome wide analyses of higher eukaryotic MITEs, FINDMITE and MUST generate a large number of false positives because many sequences that satisfy the defined parameters can occur by chance. MITE-Hunter was developed to decrease the number of false positives [30]. It uses multiple sequence alignment to filter out sequences otherwise meeting MITE signature criteria but bearing similar flanking sequences. As a result, MITE-Hunter has a false positive rate of 4.4-8.3% compared to 85% of FINDMITE and 86% of MUST [30]. In these programs, all of the candidate elements were retrieved and analyzed while, theoretically, the identification of only a single element is necessary for a MITE family with hundreds of copies. Therefore the existing algorithms are resource expensive and require lengthy processing time. For example, it took MITE-Hunter ~44 hours to process the rice genome database (~380 Mb) with a Linux cluster using five CPUs. The 700 raw output entries were reduced and grouped into 132 MITE families with manual downstream analyses.

Here, an algorithm was developed to increase processing efficiency by reducing or avoiding redundant computing, therefore shortening the processing time and reducing the requirement for computing resources. The novel algorithm was implemented in MITE Digger, one of the few TE analysis programs featuring graphical user interface [31,32]. When tested with the rice genome sequence database, it took MITE Digger ~15 hours on a quad core Windows system to complete processing with a typical memory use of ~150 Mb. Comparative analyses of the MITE Digger output with the MITE Hunter output showed that MITE Digger is accurate with low false positive and false negative rates.

Methods

Database and programs

The rice genome sequence database was obtained from IRGSP/RAP build 5 [33]. The output from MITE-hunter was obtained from Yujun Han and Sue Wessler (personal communication) [30]. Analyses of MITE families and comparisons were performed with MAK1.8 [34,35] (http://labs.csb.utoronto.ca/yang/MAK/). MITE Digger was based on the Perl script used for the retrieval of ATon elements [36]. Blast + 2.2.24 was used to perform sequence similarity searches. MITE Digger output was generated using its default parameters. The MITE Digger entries that do not match those in the MITE-hunter output were searched against the rice repeat database using RepeatMasker (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker). Genome sequences of rapa, tomato, potato and sorghum were downloaded from plantGDB (http://www.plantgdb.org/).

Implementation and testing system

The algorithm was implemented in ActivePerl 5.10.1 with Perl Tk 804.029 and Bioperl modules [37]. MITE Digger was tested on Windows XP and Windows 7 systems.

Results and discussion

Redundant computing in genome wide discovery of MITEs

A MITE family typically consists of several hundred highly similar copies. Therefore, when every candidate element is processed in genome wide analyses, a MITE family can be computed hundreds of times. Such repetitions can occur at multiple stages including signature feature (i.e. TIR and TSD) identification, screening, multiple sequence alignment or clustering. These repetitions also occur to the elements that do not qualify for the input criteria such as the element length. Furthermore, genomes often contain highly repetitive sequences such as retro elements that have the structures of short inverted repeats flanked by short direct repeats buried in their internal sequences. These non-MITE sequences can occur more often than MITE families and each family may be computed hundreds or thousands of times depending on their copy numbers.

These redundant computing can take up a major portion of the total processing time. In addition, after retrieval of every element of a family, the need to remove redundancy or to group elements into a family can be time consuming and resource intensive. Particularly, when multiple sequence alignment is used to identify candidates with different flanking sequences, aligning elements in a family with several hundred copies takes a significant amount of time. Finally, when database entries are sliced into very short fragments to reduce memory use, the processing efficiency can be dramatically affected because of the overhead on retrieving and analyzing a large number of sequences.

Measures to reduce redundant computing

Elimination or reduction of these redundant computing should dramatically increase efficiency. Theoretically, only a single complete copy of a MITE family is sufficient to represent the whole family and repetitions on other copies are unnecessary. Therefore, computing on each member of a candidate MITE family or a false family can be reduced to the computing on a representative of the family. Once a representative is identified, later occurrences of the family can be filtered out from the database entries. The redundancy is heavily concentrated toward the later part of a genome sequence database. If the members of a family with N copies are distributed randomly in a genome, the probability (P) of not having any copy present in the first X portion of the genome is: $P = (1\text{-}X)^N$, therefore the portion of the genome to have at least one copy at this probability can be calculate as $X = 1\text{-}P^{1/N}$. The portion of a genome to have at least one copy of a family with 10, 20, 50 copies at different probability level can be calculated (Figure 1A). Accordingly, the probability of missing an element of a MITE family with 10, 20, or 50 copies in 37%, 21% and 8% of a genome database is only 0.01, i.e. the probability to contain at least one copy of these families in respective portions of the genome database is high ($P = 0.99$). Less than 5% of the genome database is needed to contain at least one copy for MITE families with >100 copies with the probability of 0.995, i.e. probability of 0.005 to miss (Figure 1B). Therefore, only a small portion of a genome may be necessary to retrieve most of the MITE families; this is particularly useful in the processing of large genome databases. The members of a MITE family are normally expected to be located in different genomic loci, a false family can be identified as that having similar flanking sequences for every individual copy. Instead of an alignment of the flanking sequences of all copies of a family, comparison of those of a small subset of the members is sufficient and can reduce processing time dramatically. In addition, the removal of low complexity regions at start can eliminate the need to process false TIR and TSD motifs that are often found in such regions. Finally, processing the database entries in large sizes with high sensitivities will reduce the overhead of processing numerous small entries.

Pipeline algorithm and parameters

MITE Digger takes a genome sequence database file in FASTA format as input. The entries are automatically sliced to the maximal size of 100,000 nt if the entries are larger than this size (Figure 2). Each entry of the sliced database was used to search against a reverse complementary sequence of the entry to reduce memory consumption and increase sensitivity. This search was performed with highly sensitive parameters (e.g. evalue 10000) to reveal very short inverted repeats. Tabular output format was used to reduce memory use. Simple inverted repeats such as those with stretches of a single or dinucleotides and those only contain G/C or A/T were ignored. A pair of short stretches of sequences (10 nt) flanking the inverted repeats within the defined length range (e.g. 50–800 nt) were used to find direct repeats of 2 to 10 nt in

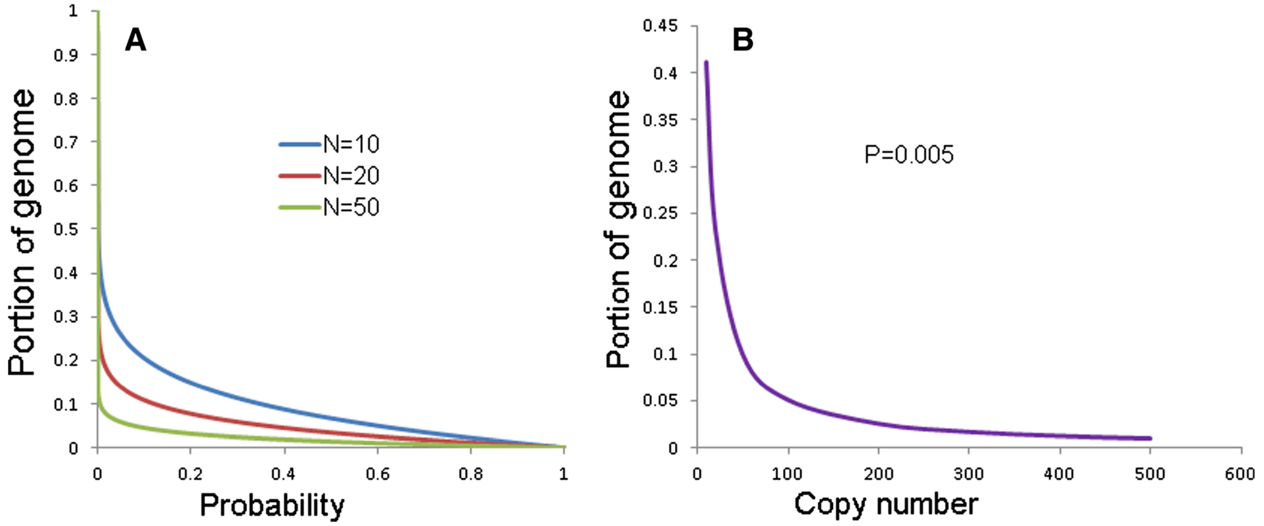

Figure 1 Analysis of redundant computing in genome wide MITE discovery. (**A**) The portions of a genome sequence database at which to miss an element of a MITE with N (10, 20, 50) copies at different probability levels; (**B**) The portion of a genome sequence database at which to miss an element with a probability of P = 0.005, i.e. P = 0.995 to contain an element, for families with different copy numbers.

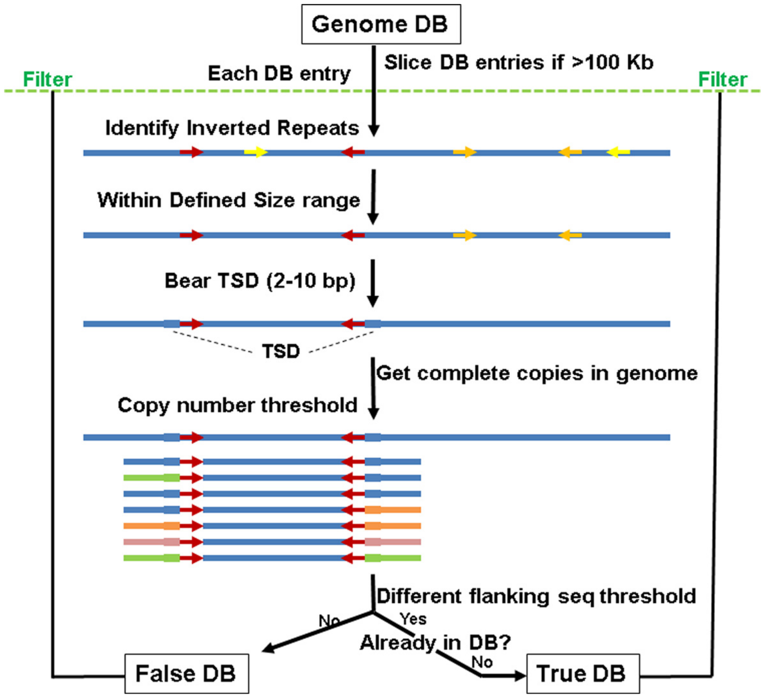
Genome DB
Filter
Each DB entry
Slice DB entries if >100 Kb
Filter
Identify Inverted Repeats
Within Defined Size range
Bear TSD (2-10 bp)
TSD
Get complete copies in genome
Copy number threshold
Different flanking seq threshold
No
Yes
Already in DB?
No
False DB
True DB

Figure 2 Algorithm of MITE digger. Arrowheads of the same color, inverted repeat pairs; Colored lines flanking inverted repeat pairs, different flanking sequences. DB, database; TSD, target site duplication.

size. Since TSDs of some elements such as *Stowaway*s and *PiggyBac* TEs may be present in inverted repeats, the search for direct repeats is also performed on the most terminal 5 nt of the inverted repeats if direct repeats are not found in the flanking sequences. The DNA sequences between each qualified inverted repeats were then used to search against the whole genome database and the number of full length copies were counted. When a candidate meets the copy number threshold, the flanking sequences of 50 nt on each side were retrieved. The flanking sequences of different copies of the candidate family were compared. When qualified for the number of different flanking sequences, the candidate sequence was deposited into the filter database and the output if a similar sequence is not already in the output. Otherwise, the sequences meeting the copy number threshold but not the number of different flanking sequences are deposited into the filter database and the false output. The filter database is used to mask the upcoming genome database entries. The processing information is printed to a report file each time a new candidate is identified. When a new candidate or false sequence is deposited, the filter database and the output database (as applicable) were reformatted for BLAST searches.

Performance analyses

MITE Digger was used to process the rice genome database to identify MITEs with a size range of 50 to 800 nt. On a Windows XP system with 4 CPU cores, processing finished in 15.44 hours with a typical memory use of ~150 Mb. A total of 332 candidate MITE sequences and 3071 false sequences were generated (Additional file 1). Entries from MITE Digger output can be used directly for downstream analyses. When the numbers of output entries were plotted against the portions of the genome processed, it is apparent that the number of candidates retrieved in a certain amount of genome sequences decreases with the increase of the amount of processed genome sequences. This reflects the reduction in redundant computing (Figure 3A). The most apparent increase in the number of candidates is in the first 20% of the genome. The curve nearly flattens out for the last 20% of the genome. While the processing rate fluctuates because of different levels of complexity of sequences in different regions, it clearly more than doubled from <3% of the genome sequence database per hour at the beginning to >6% at the end of the processing (Figure 3B). Predicted remaining time to completion decreases dramatically particularly during the first 10% of the genome database. Filtering has also dramatically reduced the number of false elements per database unit over the course (Figure 3C).

MITE Digger allows customized input of parameters. Database entries larger than the defined entry size will be automatically sliced and automatically formatted. The option to set the number of CPUs allows optimal performance of MITE Digger in platforms with different

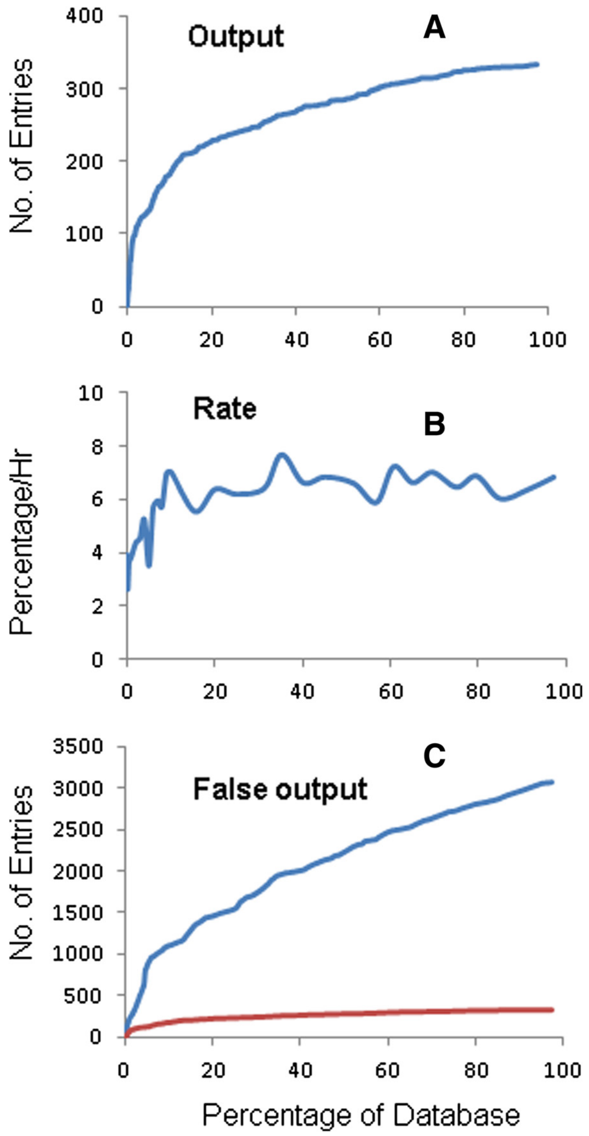

Figure 3 Processing of the rice genome by MITE digger. (**A**) The number of entries in the output database at certain processed portion of the database; (**B**) The processing rate, expressed as the percentage of the database processed per hour, at certain database percentage; (**C**) The number of entries in the false output database at certain processed portion of the database. Red, the number of entries in the output shown in (**A**).

hardware settings. The option for probability level allows timely processing of large genome databases with a minor chance of missing a MITE family. Changes to other parameters such as the copy number threshold, different flank sequence threshold and sensitivity will affect the number of output entries. The predicted running time is based on the current average processing rate, therefore the actual total run time can be dramatically shorter than the predicted time at the early part of processing because of acceleration (Figure 3B).

Evaluation of MITE Digger output

The output from MITE Digger was compared with that of the MITE Hunter. First, the entries in the MITE Digger output were cross matched with those in the MITE Hunter output, resulting in 1407 non-redundant matching pair records. Since the MITE Hunter output contains entries up to 1500 bp, only those between 50 and 800 bp were considered as MITEs. Among the 1407 records, 658 pairs cover at least 80% of the length of both query and hit sequences. The remaining records were manually inspected and four additional matching pairs were found. Because a MITE family may contain several subfamilies, one entry from MITE Digger can match several entries in MITE Hunter and *vice versa*. Therefore, the 662 matching pairs consisted of only 287 MITE Digger entries and 301 MITE Hunter entries (Additional file 2). In the remaining 749 records, 13 MITE Digger entries match the terminal sequences of some MITE Hunter entries, suggesting that these MITE Digger entries represent new subfamilies in the families of the MITE Hunter entries (Additional file 2). The rest of the records match short regions in the internal sequences of some long entries of the MITE Hunter output and were considered non-informative records. Therefore, of the 332 MITE Digger output entries, 300 can be classified with the MITE Hunter output entries. The remaining 32 entries were scanned with rice repeat database using RepeatMasker, 11 of them were previously annotated elements or can be manually classified. Six (1.8%) of them are false positives: two rice simple repeats and four retroelements (one SZ-50_int-int LTR terminal, two Cassandra, and one SINE03_OS). The other five were DNA elements: two *Stowaway*s (TREP215, STOWAWAY10_OS); one *Harbinger* (ID-4), one *Mutator* and one unknown category (OSTE23). Therefore a total of 17 (7.5%) MITE Digger output entries are classifiable MITE families that are not in the MITE Hunter output (Additional file 2). The remaining 21 entries cannot be classified even though they have the characteristics of Class II TEs.

To estimate the accuracy of MITE Digger, the output was compared with the MITE entries in Repbase and the entries in MITE Hunter output. Repbase contains 165 entries of non-autonomous Class II elements shorter than 800 bp. Among these sequences, only 19 did not have a match in the MITE Digger output. None of these elements meet the criteria for MITE: five (CASMALL, F804, HELIA, POP-OL2, WUJI) do not bear TIRs and the other 14 (CASIN, ECSR, GLUTEL1LIKE, HEARTBLEEDING, ID-2, NONAME, OSTE26, STONE, TOURIST-XIII, CACTA-G1, LIER, TAMI2, THRIA, TYPEU3) have less than 10 complete copies in the genome. Therefore, all of the MITE entries in Repbase were retrieved by MITE Digger. The accuracy was also calculated as the percentage of the MITE Digger output entries matching entries

Table 1 Summary of the comparison between the MITE digger and MITE hunter output

Category	MITE hunter output	MITE digger output	Matching entries	Only in hunter	Only in digger
<= 800 bp	462	332	(Digger-Hunter)		
MITE (<=800 bp, >10 copies)	239/462	332	287-301		
CACTA	2/3	1	1-2		
mhAT	47/102	49	48-60	5	1
mMariner/Stowaway	64/95	91	88-70	6	3
mMutator	58/134	78	70-76	7	8
mPIFHarbinger	68/128	85	80-93		5
<10 copies	223				
>800 bp	75				
False positive (retro, simple)	58/700 (8.3%)	6/332 (1.8%)			
False negative		3/332 (0.9%)			
Unclassified	15/700	22/332 (6.6%)			
Compound/Fusion	26/700				
Redundant	50/700				
Missing%				=18/239 (7.5%)	=17/226 (7.5%)
Total	700	332			

of the MITE Hunter output in the same size range (i.e. <=800 bp and > =10 copies in this case). Of the MITE Hunter entries, 462 are shorter than 800 bp and 239 of them also have at least 10 full length copies. Of these 239 MITE Hunter entries, 221 entries were found in the MITE Digger output, resulting in a missing rate for MITE Digger of 7.5% (18/239) (Table 1). Close inspection of these 18 elements revealed that, with the exception of OS_mhAT99 and OS_mhAT23, all of them have a significant part (50 – 242 bp, or 9%-87%) of their sequences matching an entry in the MITE Digger output (Table 1), suggesting that these elements were missed because elements of a different subfamily are present in the MITE Digger output. The reason OS_mhAT99 and OS_mhAT23 were missed remains unknown.

To calculate the false negative rate, the entries of the MITE Digger false output were cross matched with the MITE Hunter output entries. A large number of matching pairs are those matching the internal regions of MITEs of the MITE Hunter output as expected. Only three MITE Digger false output entries were found to be real MITEs that matched MITE Hunter output entries (OS_mMutator_126, OS_mMutator_69, OS_mMutator_67), therefore the false negative rate is 0.9% (3/332).

To see how fast MITE Digger runs with other genomes, the recently released genome databases of banana (*Musa acuminate*), potato (*Solanum tuberosome*), tomato (*Solanum lycopersicum*), Chinese cabbage (*Brassica rapa*) and sorghum (*Sorghum bicolor*) were used. Genome wide characterization of MITE families has not been previously performed in these genomes. All of these genomes gave much fewer numbers of entries than rice in the MITE output (Table 2). It is apparent that the running times for the databases are not simply determined by the size of a database. The factors affecting running time include the size and complexity of a genome, the number of entries in a database, the number of candidate and false entries.

Conclusions

In summary, MITE Digger retrieved exemplars of MITE families from the rice genome with high accuracy and low false positive and false negative rates. Importantly, MITE Digger is not computing resource intensive and the output requires minimal manual processing. Therefore, it can be

Table 2 MITE digger processing of additional genome databases

Genome	Genome DB size (Mb)	Processing time	No. output entries	No. DB entries	No. false output entries
Brassica rapa	268	19.76	85	42919	4111
Solanum tuberosome	567	37.06	134	5832	6997
Solanum lycopersicum	716	57.97	113	7425	9134
Sorghum bicolor	652	21.35	227	6267	3946
Oryza sativa	388	15.44	332	3734	3071
Musa acuminata	383	12.57	28	3930	1992

used routinely to perform genome wide identification of MITEs in higher eukaryotic genomes.

Availability and requirements

Project Name: MITE Digger
Project homepage: http://labs.csb.utoronto.ca/yang/MITEDigger
Operating system: Windows
Programming language: PERL
Other requirements: N/A
License: by the developer
Any restrictions to use by non-academics: license needed

Additional files

Additional file 1: MITE Digger report and output for the rice genome database.

Additional file 2: Comparison of the output of MITE Digger and MITE Hunter.

Abbreviations

TE: Transposable elements; MITE: Miniature inverted repeat transposable element; TIR: Terminal inverted repeat; TSD: Target site duplication.

Competing interests

The author declares that they have no competing interests.

Authors' contributions

GY developed and implemented the algorithm, performed tests and analyses, wrote the paper.

Acknowledgment

The MITE Hunter output was kindly provided by Dr. Yujun Han and Dr. Susan Wessler. I would like to thank Dr. Brad Cavinder for comments on the manuscript. Funded by National Sciences and Engineering Research Council (RGPIN371565 to G.Y.) of Canada; Canadian Foundation for Innovation (24456 to G.Y.); Ontario Research Fund (24456 to G.Y.) and University of Toronto.

Received: 11 March 2013 Accepted: 2 June 2013
Published: 7 June 2013

References

1. Isam F, Rooke R, Wong A, Hui C, Luu T, Bhardwaj P, Yang G: **Miniature Inverted-Repeat Transposable Elements (MITEs): Discovery, Distribution and Activity.** *Genome* 2013. doi:10.1139/gen-2012-0174.
2. Feschotte C, Zhang X, Wessler SR: **Miniature Inverted-Repeat Transposable Elements (MITEs) and their relationship with established DNA transposons.** In *Mobile DNA II*. Edited by Craig N, Craigie R, Gellert M, Lambowitz A. Washington DC: American Society of Microbiology Press; 2002:1147–1158.
3. Jiang N, Feschotte C, Zhang XY, Wessler SR: **Using rice to understand the origin and amplification of Miniature Inverted Repeat Transposable Elements (MITEs).** *Curr Opin Plant Biol* 2004, **7**(2):115–119.
4. Yang G, Zhang F, Hancock CN, Wessler SR: **Transposition of the rice miniature inverted repeat transposable element mPing in Arabidopsis Thaliana.** *Proc Natl Acad Sci USA* 2007, **104**(26):10962–10967.
5. Yang GJ, Nagel DH, Feschotte C, Hancock CN, Wessler SR: **Tuned for transposition: molecular determinants underlying the hyperactivity of a stowaway MITE.** *Science* 2009, **325**(5946):1391–1394.
6. Lu C, Chen JJ, Zhang Y, Hu Q, Su WQ, Kuang HH: **Miniature Inverted-Repeat Transposable Elements (MITEs) have been accumulated through amplification bursts and play important roles in gene expression and species diversity in Oryza sativa.** *Mol Biol Evol* 2012, **29**(3):1005–1017.
7. Yaakov B, Ben-David S, Kashkush K: **Genome-wide analysis of stowaway-like MITEs in wheat reveals high sequence conservation, gene association, and genomic diversification.** *Plant Physiol* 2013, **161**(1):486–496.
8. Park KC, Lee JK, Kim NH, Shin YB, Lee JH, Kim NS: **Genetic variation in Oryza species detected by MITE-AFLP.** *Genes Genet Syst* 2003, **78**(3):235–243.
9. Momose M, Abe Y, Ozeki Y: **Miniature inverted-repeat transposable elements of stowaway are active in potato.** *Genetics* 2010, **186**(1):59–U115.
10. Jiang N, Bao ZR, Zhang XY, Hirochika H, Eddy SR, McCouch SR, Wessler SR: **An active DNA transposon family in rice.** *Nature* 2003, **421**(6919):163–167.
11. Yano M, Katayose Y, Ashikari M, Yamanouchi U, Monna L, Fuse T, Baba T, Yamamoto K, Umehara Y, Nagamura Y, *et al*: **Hd1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the arabidopsis flowering time gene CONSTANS.** *Plant Cell* 2000, **12**(12):2473–2483.
12. Yang GJ, Lee YH, Jiang YM, Shi XY, Kertbundit S, Hall TC: **A two-edged role for the transposable element Kiddo in the rice ubiquitin2 promoter.** *Plant Cell* 2005, **17**(5):1559–1568.
13. Yan YS, Zhang YM, Yang K, Sun ZX, Fu YP, Chen XY, Fang RX: **Small RNAs from MITE-derived stem-loop precursors regulate abscisic acid signaling and abiotic stress responses in rice.** *Plant J* 2011, **65**(5):820–828.
14. Kuang HH, Padmanabhan C, Li F, Kamei A, Bhaskar PB, Shu OY, Jiang JM, Buell CR, Baker B: **Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: new functional implications for MITEs.** *Genome Res* 2009, **19**(1):42–56.
15. Cantu D, Vanzetti LS, Sumner A, Dubcovsky M, Matvienko M, Distelfeld A, Michelmore RW, Dubcovsky J: **Small RNAs, DNA methylation and transposable elements in wheat.** *BMC Genomics* 2010, **11:**408. doi:10.1186/1471-2164-11-408.
16. Piriyapongsa J, Jordan IK: **A Family of human MicroRNA genes from Miniature Inverted-Repeat Transposable Elements.** *PLoS One* 2007, **2**(2):e203. doi:10.1371/journal.pone.0000203.
17. Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, Okumoto Y, Tanisaka T, Wessler SR: **Unexpected consequences of a sudden and massive transposon amplification on rice gene expression.** *Nature* 2009, **461**(7267):1130–U1232.
18. Fattash I, Bhardwaj P, Hui C, Yang G: **A rice stowaway MITE for gene transfer in yeast.** *PLoS One* 2013. In press.
19. Yang G, Weil CF, Wessler SR: **A rice Tc1/mariner-like element transposes in yeast.** *Plant Cell* 2006, **18**(10):2469–2478.
20. Wang S, Zhang LL, Meyer E, Bao ZM: **Genome-wide analysis of transposable elements and tandem repeats in the compact placozoan genome.** *Biol Direct* 2010, **5:**520. doi:10.1186/1471-2164-11-520.
21. Feschotte C, Swamy L, Wessler SR: **Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (MITEs).** *Genetics* 2003, **163**(2):747–758.
22. Han MJ, Shen YH, Gao YH, Chen LY, Xiang ZH, Zhang Z: **Burst expansion, distribution and diversification of MITEs in the silkworm genome.** *BMC Genomics* 2010, **11:**520. doi:10.1186/1471-2164-11-520.
23. Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, Loftus B, Xi Z, Megy K, Grabherr M, *et al*: **Genome sequence of Aedes aegypti, a major arbovirus vector.** *Science* 2007, **316**(5832):1718–1723.
24. Bureau TE, Wessler SR: **Tourist: a large family of small inverted repeat elements frequently associated with maize genes.** *Plant Cell* 1992, **4**(10):1283–1294.
25. Bureau TE, Wessler SR: **Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants.** *Plant Cell* 1994, **6**(6):907–916.
26. Bergman CM, Quesneville H: **Discovering and detecting transposable elements in genome sequences.** *Brief Bioinform* 2007, **8**(6):382–392.
27. Lerat E: **Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs.** *Heredity* 2010, **104**(6):520–533.
28. Tu ZJ: **Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, Anopheles gambiae.** *Proc Natl Acad Sci USA* 2001, **98**(4):1699–1704.
29. Chen Y, Zhou F, Li G, Xu Y: **MUST: a system for identification of miniature inverted-repeat transposable elements and applications to Anabaena variabilis and Haloquadratum walsbyi.** *Gene* 2009, **436**(1–2):1–7.
30. Han Y, Wessler SR: **MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences.** *Nucleic Acids Res* 2010, **38**(22):e199.

31. Rooke R, Yang GJ: **TE Displayer for post-genomic analysis of transposable elements.** *Bioinformatics* 2011, **27**(2):286–287.
32. Tempel S, Jurka M, Jurka J: **VisualRepbase: an interface for the study of occurrences of transposable element families.** *BMC Bioinformatics* 2008, **9**:345. doi:10.1186/1471-2105-9-345.
33. Tanaka T, Antonio BA, Kikuchi S, Matsumoto T, Nagamura Y, Numa H, Sakai H, Wu J, Itoh T, Sasaki T, *et al*: **The rice annotation project database (RAP-DB): 2008 update.** *Nucleic Acids Res* 2008, **36**:D1028–D1033.
34. Yang GJ, Hall TC: **MAK, a computational tool kit for automated MITE analysis.** *Nucleic Acids Res* 2003, **31**(13):3659–3665.
35. Janicki M, Rooke R, Yang GJ: **Bioinformatics and genomic analysis of transposable elements in eukaryotic genomes.** *Chromosome Res* 2011, **19**(6):787–808.
36. Yang GJ, Wong A, Rooke R: **ATon, abundant novel nonautonomous mobile genetic elements in yellow fever mosquito (Aedes aegypti).** *BMC Genomics* 2012, **13**:283. doi:10.1186/1471-2164-13-283.
37. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, *et al*: **The Bioperl toolkit: perl modules for the life sciences.** *Genome Res* 2002, **12**(10):1611–1618.

doi:10.1186/1471-2105-14-186
Cite this article as: Yang: MITE Digger, an efficient and accurate algorithm for genome wide discovery of miniature inverted repeat transposable elements. *BMC Bioinformatics* 2013 **14**:186.