Ruiquan Ge[1,a] / Guoqin Mai[2,a] / Ruochi Zhang[3,a] / Xundong Wu[1] / Qing Wu[1] / Fengfeng Zhou[3]

MUSTv2: An Improved *De Novo* Detection Program for Recently Active Miniature Inverted Repeat Transposable Elements (MITEs)

[1] School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, Zhejiang 310018, China, E-mail: wuq@hdu.edu.cn

[2] Center for Synthetic Biology Engineering Research, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, China

[3] College of Computer Science and Technology, and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin 130012, China, E-mail: ffzhou@jlu.edu.cn

Abstract:

Background **:** Miniature inverted repeat transposable element (MITE) is a short transposable element, carrying no protein-coding regions. However, its high proliferation rate and sequence-specific insertion preference renders it as a good genetic tool for both natural evolution and experimental insertion mutagenesis. Recently active MITE copies are those with clear signals of Terminal Inverted Repeats (TIRs) and Direct Repeats (DRs), and are recently translocated into their current sites. Their proliferation ability renders them good candidates for the investigation of genomic evolution.

Results **:** This study optimizes the C++ code and running pipeline of the MITE Uncovering SysTem (MUST) by assuming no prior knowledge of MITEs required from the users, and the current version, MUSTv2, shows significantly increased detection accuracy for recently active MITEs, compared with similar programs. The running speed is also significantly increased compared with MUSTv1. We prepared a benchmark dataset, the simulated genome with 150 MITE copies for researchers who may be of interest.

Conclusions **:** MUSTv2 represents an accurate detection program of recently active MITE copies, which is complementary to the existing template-based MITE mapping programs. We believe that the release of MUSTv2 will greatly facilitate the genome annotation and structural analysis of the bioOMIC big data researchers.

Keywords: miniature inverted repeat transposable element (MITE), recently active MITE copy, MUST program, terminal inverted repeat, direct repeat

DOI: 10.1515/jib-2017-0029

Received: April 6, 2017; **Revised:** April 20, 2017; **Accepted:** May 8, 2017

1 Background

Transposable elements are genetic components that trans-locate themselves within or across genomes, and are classified as autonomous or non-autonomous elements based on whether they encode the enzymes for translocations [1], [2]. Miniature inverted repeat transposable elements (MITEs) have been found in both eukaryota and prokaryota [1], [3]. MITEs and other transposons played essential roles in the evolution procedure of plant genomes [4], [5], [6]. Although MITEs are short in lengths (100–600 bps), they are well-known for their active translocations and high copy numbers [3], [7]. This feature makes them ideal for developing insertion mutagenesis techniques.

MITEs have a number of sequence level features. They are usually 100–600 bps in length, have a pair of highly precise reversely complementary terminal inverted repeats (TIRs), are flanked by a pair of almost identical direct repeat (DRs), and tend to be AT-rich [3], [8]. Unfortunately, the majority of large-scale transposable element (TE) annotation studies focus on the sequence similarity of TE copies to the template, and provide no knowledge about MITE's sequence features, i.e. TIRs and DRs [9]. A recently active MITE copy has a pair of TIRs and DRs, and is almost identical to its parent copy. The host genome evolution facilitated by the natural insertion mutagenesis only occurs when MITEs proliferate. So it is important to detect the recently active

Qing Wu, Fengfeng Zhou are the corresponding authors.

[a] Ruiquan Ge, Guoqin Mai and Ruochi Zhang: These authors contribute equally to this work.

MITE copies and to investigate how they proliferate within and between genomes. Some species-specific MITE databases were curated from the literature [10], [11], [12].

A number of *de novo* MITE detection programs have been released, and they screen a given genome for all the copies of candidate MITEs [13], [14]. FINDMITE is one of earliest *de novo* MITE detection programs, but its downloading web site is not available now [15]. MITE-Digger [16] performs faster, but is less sensitive, compared with the another program, MITE-Hunter [17]. Both programs do not provide the TIRs and DRs of each MITE copy in the final annotations. MITE-Digger works only in Windows, whereas MITE-Hunter is independent of computer operating systems. MUST version 1.0 (MUSTv1) is highly sensitive in detecting novel MITEs, but is also high in false positive rates [7], [17].

We collect the comments from the MUSTv1 external users and our collaborators, and significantly restructure the MUST detection algorithm. The current version MUSTv2 has extremely high accuracy, demonstrated using a simulated genome with 150 inserted MITE copies and the rice genome. The running speed of MUSTv2 is almost doubled for the processing of large genomes, per the user suggestions.

2 Implementation

2.1 Versions and Parameters of Programs Used in MUSTv2

This study used the “open-3.3.0” version of RepeatMasker, with the revision 1.250 on April 26, 2011. The software was updated and tested in our computing server on March 15, 2017. The most widely used alignment software, NCBI BLASTALL version 2.2.11, was chosen as the match searching engine of RepeatMasker. BLAT version 3.5 was used in MUSTv2.

MITE-Hunter has no version information, and the downloaded version of MITE-Hunter has a last updating record on August 19, 2010. This MITE-Hunter package was installed in the computing server on October 20, 2011.

The version 2.4.002 of MUSTv2 was used in this study.

All the programs were executed with their default parameters.

2.2 Detection Procedure of MUSTv2

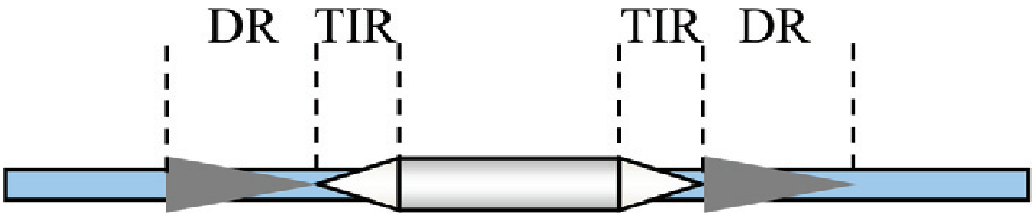


Figure 1: Structure of a MITE. A MITE has a pair of terminal inverted repeats (TIRs) in the boundary and a pair of direct repeats (DRs) in the direct flanking region.

MUSTv2 implements the following procedure to detect MITEs in a given genome, and all the parameters may be changed in the command line. The structure of a MITE is illustrated in Figure 1.

1. Detect all the pairs of TIRs satisfying the given parameters in the given genome, with the following parameters

a. (MinTIR, MaxTIR): minimum and maximum lengths of TIRs, with defaults 8 and 50.

b. (MinDR, MaxDR): minimum and maximum lengths of DRs, with defaults 2 and 30.

c. (MinMITE, MaxMITE): minimum and maximum lengths of MITEs, with defaults 100 and 600.

d. FixedFlanking: the length of flanking regions for screening DRs, which must be greater than MaxDR and has the default 50.

e. MutationRate: variations between pairs of TIRs and DRs, with default 0.80.

2. Test whether a given pair of TIRs is flanked by a pair of DRs.

a. This pair of TIRs will be skipped, if not satisfying this testing.

3. Cluster the candidate MITEs, and only keep those clusters with minimally required copy numbers, and the same TIR/DR signals.

a. This is to group copies of the same MITE together

4. Due to the possible mutations in TIRs and DRs, another round of screening for all the full copies of the detected MITE copies is conducted.

a. To confirm whether the MITE copies are full copies, clustered from the above step.
5. Re-screen the newly detected copies for the same TIR/DR signals of the same cluster.
a. To re-detect the TIRs and DRs of the confirmed full copies, whose boundaries may be slightly changed.
6. Summarize and output all the detected MITE copies, together with their detailed sequence structures.
a. Summarize the current comprehensive annotations into human-readable format.

There are three major improvements for the MUST algorithm. Firstly, MUSTv1 used a slow Markov Chain cLustering (MCL) algorithm [18] to group the detected candidate MITEs into families, and MUSTv2 implements a heuristic clustering algorithm by joining a given element with the element of the largest overlap in the same group. Secondly, by focusing on the recently proliferated MITE copies, MUSTv2 now requires the external boundaries of TIRs to be strictly reversely complementary and only allows mutations in the TIR internal regions, whereas MUSTv1 allows mutations in any positions of TIRs. Thirdly, all the for-loops in the code are optimized to increase the running speeds.

MUSTv2 is implemented using Perl/C++ program languages, and BioPerl library [19] is also used to process the sequence files. This setting makes MUSTv2 portable to any Linux/Unix-based high performance computing environments and Windows/Linux/Unix/AppleOS-based personal computers. This study is conducted in a Linux-based computing server with 48 Gb memory and 14 Tb hard disk.

3 Results and Discussion

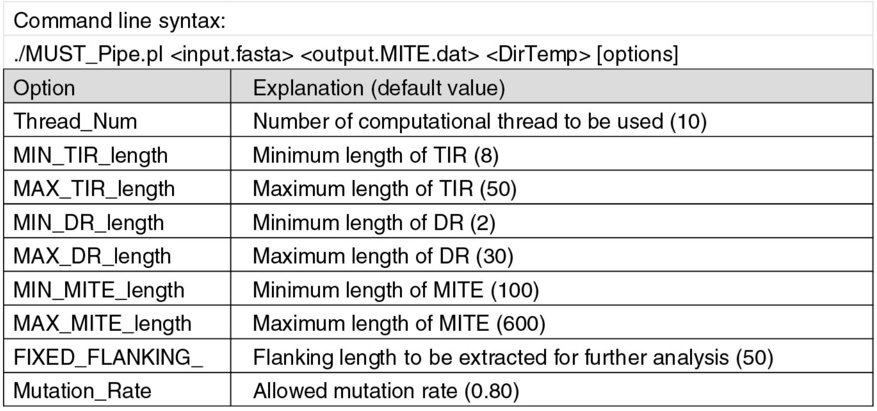
3.1 The Reference Data

There is no gold standard dataset of MITEs, and this study chooses to generate a simulated genome with inserted known MITE copies. The chromosome of *Escherichia coli* K12 MG1655 is chosen as the host genome. Three MITEs with clear sequence features are used as the template MITEs.

The copy gi:22830894 of *mPing* with its DR TAA is chosen [20]. The second template MITE is the copy of *Chunjie* at NC_009483: 2632106-2632324, and its DR is ACGACCGGT [8]. The third template MITE *Nezha* is retrieved from NC_007413: 2978180-2978317 with the DR CATTATCTAC [3]. Fifty copies of each template MITE are randomly inserted into the host genome, and no copy is inserted in the other MITE copies. For the ease of future comparison by other groups, we provided the chromosome sequence of *Escherichia coli* K12 MG1655 before and after the simulated insertions in FASTA format as two supplementary files and the simulated insertion sites in Supplementary Table S1.

The command line syntax was given in Figure 2.

3.2 Command Line Syntax



Command line syntax:

```
./MUST_Pipe.pl <input.fasta> <output.MITE.dat> <DirTemp> [options]
```

Option	Explanation (default value)
Thread_Num	Number of computational thread to be used (10)
MIN_TIR_length	Minimum length of TIR (8)
MAX_TIR_length	Maximum length of TIR (50)
MIN_DR_length	Minimum length of DR (2)
MAX_DR_length	Maximum length of DR (30)
MIN_MITE_length	Minimum length of MITE (100)
MAX_MITE_length	Maximum length of MITE (600)
FIXED_FLANKING_	Flanking length to be extracted for further analysis (50)
Mutation_Rate	Allowed mutation rate (0.80)

Figure 2: Command line syntax of MUSTv2. The names of the input and output files and the directory name for temporary files are required parameters. All the other parameters are optional, and their default values will be used if not being input. The meanings of the parameters are explained in the table and the above section "Detection Procedure of MUSTv2". MUSTv2 is a command line program, and no graphical user interface was provided.

3.3 Recovering MITEs in the Simulated Genome

A comparison of *de novo* MITE detection was conducted for the predictions of three programs, i.e. MUSTv1 [7], MITE-Hunter [17], and MUSTv2 (this study). MITE-Digger performs faster but detects fewer MITEs, compared with MITE-Hunter [16]. So MITE-Hunter is chosen for comparison in this study. The majority of MITE prediction programs do not generate the TIRs and DRs for each MITE copy, including MITE-Hunter. So a MITE copy is defined to be recovered, if 90% region of this MITE copy is within a predicted copy, and the signals of TIRs and DRs are only compared between MUSTv1 and MUSTv2. All the programs are executed using default parameters.

All the three programs successfully detect all the 150 MITE copies, as shown by the column SimG of Table 1. The program MUSTv1 detects 489 MITE copies. Besides the 150 simulated MITE copies, the other predicted copies also have the signals of TIRs and DRs. We cannot reject the hypothesis that these extra copies are within the segmental duplication regions, e.g. both the DRs and the flanking regions of the three copies of candidate MITE (Cluster:22) are identical to each other. For a strict definition of the detection performance, these MITEs other than the 150 simulated ones were regarded as false positives. After filtering the low complexity regions and putative IS elements, MITE-Hunter correctly detects exactly 50 copies of the three MITEs as MiteHunter | 1404_227 (length:222), MiteHunter | 241_231 (length:430) and MiteHunter | 1369_6 (length:135). But the annotations for *Chunjie* and *Nezha* are different in the boundaries of MiteHunter | 1404_227 and MiteHunter | 1369_6. Annotation MiteHunter | 589_108 is 1195 bps in length, longer than known MITEs (600 bps). MiteHunter | 589_108 overlaps with protein-coding genes, and the full copies of the other MITEs do not have the detectable signals of TIRs and/or DRs. Although MITE-Hunter claims to detect repeats as long as 2000 bps, the repeats other than the 150 simulated MITE copies detected by MITE-Hunter may not be real MITEs.

Table 1: Result summary of the three programs on the simulated genome and the rice genome.

	SimG	FDR (%)	SimG Time (s)	Rice genome		Time (h)
				Stowaway1_OS	TREP215	
Real	150	–	–	2757	2687	–
MUSTv1	150/489	69.33	1148.40	–	–	–
MUSTv2	150/150	0.00	82.35	573	655	50.42
Mite Hunter	150/450	66.67	458.20	106	13	35.23

Column SimG gives two numbers "num1/num2", which are the numbers of the recovered real copies and all the predictions, respectively. Only the full copies of MITE-Hunter predictions are counted. MUSTv1 ran 10 times slower than MUSTv2, and was terminated before finishing. Column FDR gives the false discovery rate. The default numbers of threads used by MUSTv2 and Mite Hunter are 10 and 5, respectively. Mite Hunter generated 6 consensus MITE templates. Column "Time (h)" gives the running times of the programs in hours.

MUSTv2 detects exactly 50 copies for each of the three MITEs with no false positives, as shown in Table 1. Except for 5 MITE copies, MUSTv2 predicts the correct regions, MITE length, and the TIR/DR signals. Among these 5 copies, one has 4-bp extensions for its boundaries, and the other four have 3-bp boundary extensions. This may be due to the perfect short 2–3 bp DRs flanking the predicted boundaries.

3.4 Comparison of MITE-Hunter and MUSTv2 on the Rice Genome

The rice genome was an well-annotated genome with repeats, and the program RepeatMasker has almost all the known rice transposons in its internal database. So the rice genome was annotated using RepeatMasker for known MITEs, and two MITEs with the top complete copy numbers (Stowaway1_OS and TREP215) were used for the comparison between MUSTv2 and MITE-Hunter, as shown in Table 1. The comparison of all the other MITEs in the rice genome may be found in Figure 3.

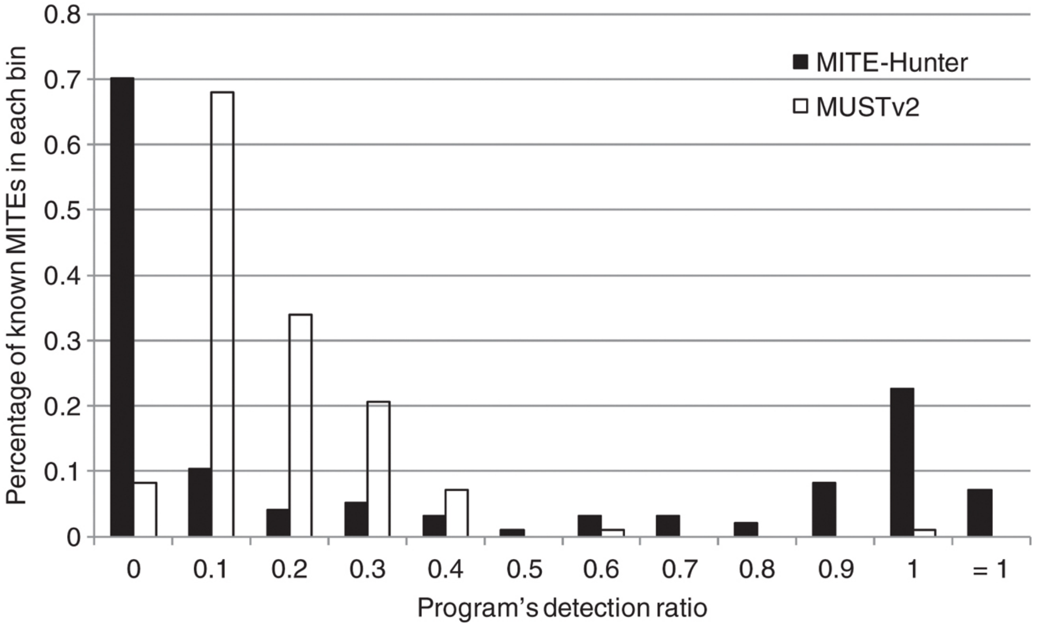


Figure 3: The histogram plots of the percentages of repeats detected by the two tools MITE-Hunter and MUSTv2, respectively. Each MITE has the percentage of full copies detected by each program, which is called the program's detection ratio of this MITE. And the Y axis is the percentage of the 136 MITEs from the Supplementary Table S2 that have the detection ratio by each of the two programs. Each bin gives the percentage of known MITEs that has the percentage of full copies detected by each program, i.e. [0, 0], (0, 0.1), [0.1, 0.2), ..., [0.9, 1.0), [1.0, 1.0].

MUSTv2 confirms that 573 of the 2757 Stowaway1_OS copies are recently active copies, since they have both clear TIR and DR signals, where MITE-Hunter only confirms 106 Stowaway1_OS copies. 655 of 2687 TREP215 copies are suggested by MUSTv2 to be structurally complete MITE copies, whereas MITE-Hunter only detects 13 copies. Due to that TREP215 has internal repetitive regions, MITE-Hunter may have split the annotations into partial copies, and significantly decreases the number of candidate full copies. Detection performance on the other elements may be found in the Supplementary Table S2 and Figure 3.

The overall specificity (or precision) of MITE calling for the two tools on the rice genome was estimated by assuming that RepBase-based RepeatMasker annotations of MITEs are comprehensive and no novel MITE exists in the rice genome. This study focuses on the MITE full copies, and the total numbers of MITE full copies detected by MUSTv2 and MITE-Hunter are 35,621 and 36,866, respectively. The full copy numbers of the 136 MITEs in the Supplementary Table S2 are 5736 and 7130, respectively. The detection specificity is defined to be the ratio between the correctly detected MITE full copy number and the total detected MITE full copy number, and this measurement is 16.10% and 19.34% for MUSTv2 and MITE-Hunter, respectively. The slightly lower detection specificity of MUSTv2 may be due to that some MITE full copies detected by MITE-Hunter lost their sequence signals, i.e. TIRs and DRs, which are required by their future proliferations and MUSTv2 detection.

Generally, MUSTv2 and MITE-Hunter performs similarly well on the detection of MITE full copies. MUSTv2 focuses on the detection the copies with complete TIRs and DRs, whereas MITE-Hunter focuses on the detection of full sequence copies.

3.5 Running Speeds

The running time of each program is calculated by the Linux command "time", and is counted in seconds. MUSTv1 runs for 1148 s for the simulated 4.68-Mbp genome, and the pipeline of MITE-Hunter and RepeatMasker runs for 458.20 s for the same genome. We have checked the detailed running steps of these two programs, and found that the screening of candidate MITEs across the whole genome, and the clustering of candidate MITEs into groups represent two major time-consuming steps. MUSTv2 optimized the MITE screening C++ codes and replaced the MCL clustering program with an in-house clustering strategy. The current version of MUST highly accurately detects all the simulated MITE copies for just 82.35 s, and achieved almost 13.95 times of the running speed of MUST system for *de novo* MITE detection. MUSTv2 runs about two times slower than MITE-Hunter on the rice genome, maybe due to that MUSTv2 screens the TIR and DR signals for all the candidate MITE copies.

3.6 Limits of MUSTv2

MUSTv2 is not good at the detection of MITEs without clear TIR and DR signals. MUSTv2 focuses on detecting the recently active MITEs, which are usually high in copy numbers and tend to have perfect pairs of TIRs and DRs. Most of the MITE copies will lose their ability of proliferation, and start accumulating point mutations and/or structural variations, after being inserted into their current locations. As discussed in the above section, it seems that the rice genome is undergoing a very rapid mutation process, and many MITE copies lost their sequence patterns like TIRs and DRs. The data suggests that both MUSTv2 and MITE Hunter did not work well on the MITE annotations in the rice genome. So the *de novo* MITE detection programs should work together with the other MITE annotation programs like RepeatMasker for a more complete and comprehensive annotation of MITEs in a given genome.

But it is the active MITE copy that keeps proliferation and applies their driving forces onto the genomic variations and evolution. So the recently active MITE copies detected by MUSTv2 will be an informative and complementary resource to the genomic MITE annotation, mainly through the template mapping techniques. MUSTv2 may also detect novel active MITEs through the *de novo* sequence structural patterns.

4 Conclusions

This study proposed a highly accurate and very fast computer program for the *de novo* detection of recently active MITE copies in a given genome. As a user-friendly program, the default parameters also make sure that the user may just give the genomic sequences in a FASTA file and the prediction result will be generated into a user-specified output file. Since there are still many unknown MITEs and other transposons in the sequenced genomes, the users may want to change the parameters to run an extensive screening for candidate novel MITEs. MITEs are known to undergo rapid mutations after their integrations into the host genomes, so the mutation rate allowed in MUSTv2 may also be decreased to detect MITE copies inserted into their current locations a long time ago.

5 Availability and Requirements

Project name: MUSTv2
Project home page: http://www.healthinformaticslab.org/supp/
Operating system(s): platform-independent, best in Linux/Unix
Programming languages: Perl, C/C++ and Bash
Other requirements: BioPerl
License: GNU GPL v2 to academic users

Acknowledgments

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13040400) and the startup grants of the Jilin University (FZ) and Hangzhou Dianzi University (RG). Comments from the two anonymous reviewers were also greatly appreciated.

Conflict of interest statement: Authors state no conflict of interest. All authors have read the journal's Publication ethics and publication malpractice statement available at the journal's website and hereby confirm that they comply with all its parts applicable to the present scientific work.

Authors' contributions: FZ conceived and oversaw this project. RG, GM and RZ developed the software and tested the results. FZ, QW and XW drafted the manuscript with the contributions from the co-authors. All the authors read and approved the final manuscript.

References

[1] Jiang N, Feschotte C, Zhang X, Wessler SR. Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). Curr Opin Plant Biol. 2004;7:115–9.
[2] Delihas N. Impact of small repeat sequences on bacterial genome evolution. Genome Biol Evol. 2011;3:959–73.
[3] Zhou F, Tran T, Xu Y. Nezha, a novel active miniature inverted-repeat transposable element in cyanobacteria. Biochem Biophys Res Commun. 2008;365:790–4.
[4] Shen J, Liu J, Xie K, Xing F, Xiong F, Xiao J, et al. Translational repression by a miniature inverted-repeat transposable element in the 3′ untranslated region. Nat Commun. 2017;8:14651.
[5] Yaakov B, Ceylan E, Domb K, Kashkush K. Marker utility of miniature inverted-repeat transposable elements for wheat biodiversity and evolution. Theor Appl Genet. 2012;124:1365–73.
[6] Konovalov FA, Goncharov NP, Goryunova S, Shaturova A, Proshlyakova T, Kudryavtsev A. Molecular markers based on LTR retrotransposons BARE-1 and Jeli uncover different strata of evolutionary relationships in diploid wheats. Mol Genet Genomics. 2010;283:551–63.
[7] 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;436:1–7.
[8] Chen Y, Zhou F, Li G, Xu Y. A recently active miniature inverted-repeat transposable element, Chunjie, inserted into an operon without disturbing the operon structure in Geobacter uraniireducens Rf4. Genetics. 2008;179:2291–7.
[9] Ragupathy R, You FM, Cloutier S. Arguments for standardizing transposable element annotation in plant genomes. Trends Plant Sci. 2013;18:367–76.
[10] Han M-J, Zhou Q-Z, Zhang H-H, Tong X, Lu C, Zhang Z, Dai F, et al. iMITEdb: the genome-wide landscape of miniature inverted-repeat transposable elements in insects. Database (Oxford). 2016;baw148. DOI: 10.1093/database/baw148.
[11] Murukarthick J, Sampath P, Lee SC, Choi BS, Senthil N, Liu S, et al. BrassicaTED - a public database for utilization of miniature transposable elements in Brassica species. BMC Res Notes. 2014;7:379.
[12] Chen J, Hu Q, Zhang Y, Lu C, Kuang H. P-MITE: a database for plant miniature inverted-repeat transposable elements. Nucleic Acids Res. 2014;42:D1176–81.
[13] Ye C, Ji G, Liang C. detectMITE: a novel approach to detect miniature inverted repeat transposable elements in genomes. Sci Rep. 2016;6:19688.
[14] Yang G, Hall TC. MAK, a computational tool kit for automated MITE analysis. Nucleic Acids Res. 2003;31:3659–65.
[15] Tu Z. Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, Anopheles gambiae. Proc Natl Acad Sci USA. 2001;98:1699–704.
[16] Yang G. MITE Digger, an efficient and accurate algorithm for genome wide discovery of miniature inverted repeat transposable elements. BMC Bioinformatics. 2013;14:186.
[17] Han Y, Wessler SR. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res. 2010;e19938.
[18] Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30:1575–84.
[19] Stajich JE. An introduction to BioPerl. Methods Mol Biol. 2007;406:535–48.
[20] Kikuchi K, Terauchi K, Wada M, Hirano HY. The plant MITE mPing is mobilized in anther culture. Nature. 2003;421:167–70.

Supplemental Material: The online version of this article offers supplementary material (DOI: https://doi.org/10.1515/jib-2017-0029).