Database, 2016, 1–7
doi: 10.1093/database/baw148
Original article

Original article

iMITEdb: the genome-wide landscape of miniature inverted-repeat transposable elements in insects

Min-Jin Han[1,†], Qiu-Zhong Zhou[2,†], Hua-Hao Zhang[3], Xiaoling Tong[1], Cheng Lu[1], Ze Zhang[2] and Fangyin Dai[1,*]

[1]State Key Laboratory of Silkworm Genome Biology, Key Laboratory for Sericulture Functional Genomics and Biotechnology of Agricultural Ministry, Southwest University, Chongqing 400715, China, [2]Laboratory of Evolutionary and Functional Genomics, School of Life Sciences, Chongqing University, Chongqing 401331, China and [3]College of Pharmacy and Life Science, Jiujiang University, Jiujiang 332000, China

*__Corresponding author:__ Tel: 0086-023-68250793; Fax: 0086-023-68251128; E-mail: fydai@swu.edu.cn

[†]These authors contributed equally to this work.

Citation details: Han,M.J., Zhou,Q.Z., Zhang,H.H. et al. iMITEdb: the genome-wide landscape of miniature inverted-repeat transposable elements in insects. *Database* (2016) Vol. 2016: article ID baw148; doi:10.1093/database/baw148

Received 13 August 2016; Revised 19 September 2016; Accepted 18 October 2016

Abstract

Miniature inverted-repeat transposable elements (MITEs) have attracted much attention due to their widespread occurrence and high copy numbers in eukaryotic genomes. However, the systematic knowledge about MITEs in insects and other animals is still lacking. In this study, we identified 6012 MITE families from 98 insect species genomes. Comparison of these MITEs with known MITEs in the NCBI non-redundant database and Repbase showed that 5701(∼95%) of 6012 MITE families are novel. The abundance of MITEs varies drastically among different insect species, and significantly correlates with genome size. In general, larger genomes contain more MITEs than small genomes. Furthermore, all identified MITEs were included in a newly constructed database (iMITEdb) (http://gene.cqu.edu.cn/iMITEdb/), which has functions such as browse, search, BLAST and download. Overall, our results not only provide insight on insect MITEs but will also improve assembly and annotation of insect genomes. More importantly, the results presented in this study will promote studies of MITEs function, evolution and application in insects.

Database URL: http://gene.cqu.edu.cn/iMITEdb/

Introduction

Miniature inverted-repeat transposable elements (MITEs) were first discovered in plants, and are widely distributed in eukaryotes (1–5). MITEs belong to class II (or DNA) transposable elements (TEs), and are non-autonomous elements derived from the internal-deletion of autonomous DNA transposons (6, 7). However, they can be mobilized by transposases encoded by their parental autonomous transposons (called trans-mobilization) or non-parental elements (called cross-mobilization) (8, 9). MITEs can be classified into different superfamilies based on the nucleotide composition of terminal inverted repeats (TIRs) and target site duplications (TSDs). Unlike other DNA transposons, MITEs often have some obvious characteristics: shorter sequence length (<800 bp), high AT content, insertion preference in or near genes and high copy numbers in a genome (10–12).

MITEs have attracted widespread attention due to their roles in gene expression, genome evolution and phenotypic diversity (13–16). MITEs not only up-regulate the expression of nearby genes by acting as new cis-regulatory elements but also down-regulate or silence the expression of some genes by small RNAs derived from these elements at the transcriptional and/or post-transcriptional levels (14, 17–20). Besides, MITEs make a great contribution on the evolution of genome size (12, 16). Furthermore, MITEs are considered as a good genetic source applied in DNA makers, transgenic vectors and effective insertion mutagen (21–24). However, most of above results were obtained from studies of plants.

Since more and more genome sequences become available, several computer programs have been developed to identify MITEs in genomes, and a larger number of MITEs have been identified in the eukaryotic genomes especially in plant genomes (13, 20, 25–28). Although several studies tried to identify MITEs in the insect genomes (3, 5, 29), the number of reported MITEs could be just the tip of the iceberg with rapidly increasing insect genome were released.

In the present study, MITEs from 98 insect genomes were identified, classified and annotated using MITE-Hunter and Repetitive Sequence with Precise Boundaries (RSPB) as well as a series of Perl scripts. We identified 6012 MITE families belonging to 16 known superfamilies in these genomes. In total 5701 of 6012 MITEs families are novel and have no matches to the previously known MITEs in the databases of Repbase and NCBI non-redundant nucleotide database. The abundance of MITEs varies greatly among the different insect species and significantly correlated with genome size. Finally, all identified MITEs are made available in a newly constructed database called iMITEdb.

Materials and methods

Data sources used in this study

Ninety-eight released insect genomes including Coleoptera (7 species), Diptera (48 species), Hemiptera (8 species), Hymenoptera (20 species), Lepidoptera (9 species), Strepsiptera (1 species), Orthoptera (1 species), Odonata (1 species), Isoptera (1 species), Thysanoptera (1 species) and Ephemeroptera (1 species) were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/) (as of 8 March 2015) (Supplementary Table S1).

Identification, classification and characterization of insect MITEs

MITE-Hunter and RSPB were used to search for MITEs in 98 insect genomes (20, 27). Briefly, the pipeline for MITEs identification included four steps (Supplementary Figure S1): (i) First, MITE-Hunter was used to search insect genomes for candidate MITEs. Then, RSPB was used to identify potential insect MITEs. In RSPB, the hunter2ref.pl script, a Perl script of RSPB, was used to skip the confirmed MITEs identified by MITE-Hunter; (ii) Each candidate MITE was used as a query in BLASTN (*e*-value $< e^{-6}$) search against the corresponding genome sequence. Candidate MITE families with copy numbers <3 were discarded. Then, multiple sequences retrieved by each candidate MITE were aligned using MUSCLE (30); (iii) Consensus sequence was generated using a Perl script, and consensus sequences >800 bp in length were discarded; (iv) Finally, the TSDs and TIRs of each MITEs were retrieved using Perl script. MITEs from each species were assigned into families through all-versus-all BLAST method. The same family was defined by nucleotide identity>80%, BLAST *e*-value $< e^{-6}$ and percent query coverage >80%. MITEs were classified into superfamily based on TIRs and TSDs (16).

Construction of insect MITEs database

A database containing the information of all insect MITEs identified in this study was constructed using Linux, PHP, Apache, MySQL and Perl as well as Common Gateway Interface.

Results and discussion

Identification, classification and abundance of MITEs in 98 insect genomes

In this study, a total of 6012 MITE families were identified in 98 insect genomes. The consensus sequences of these MITE families were used as queries in BLASTN (*e*-value <

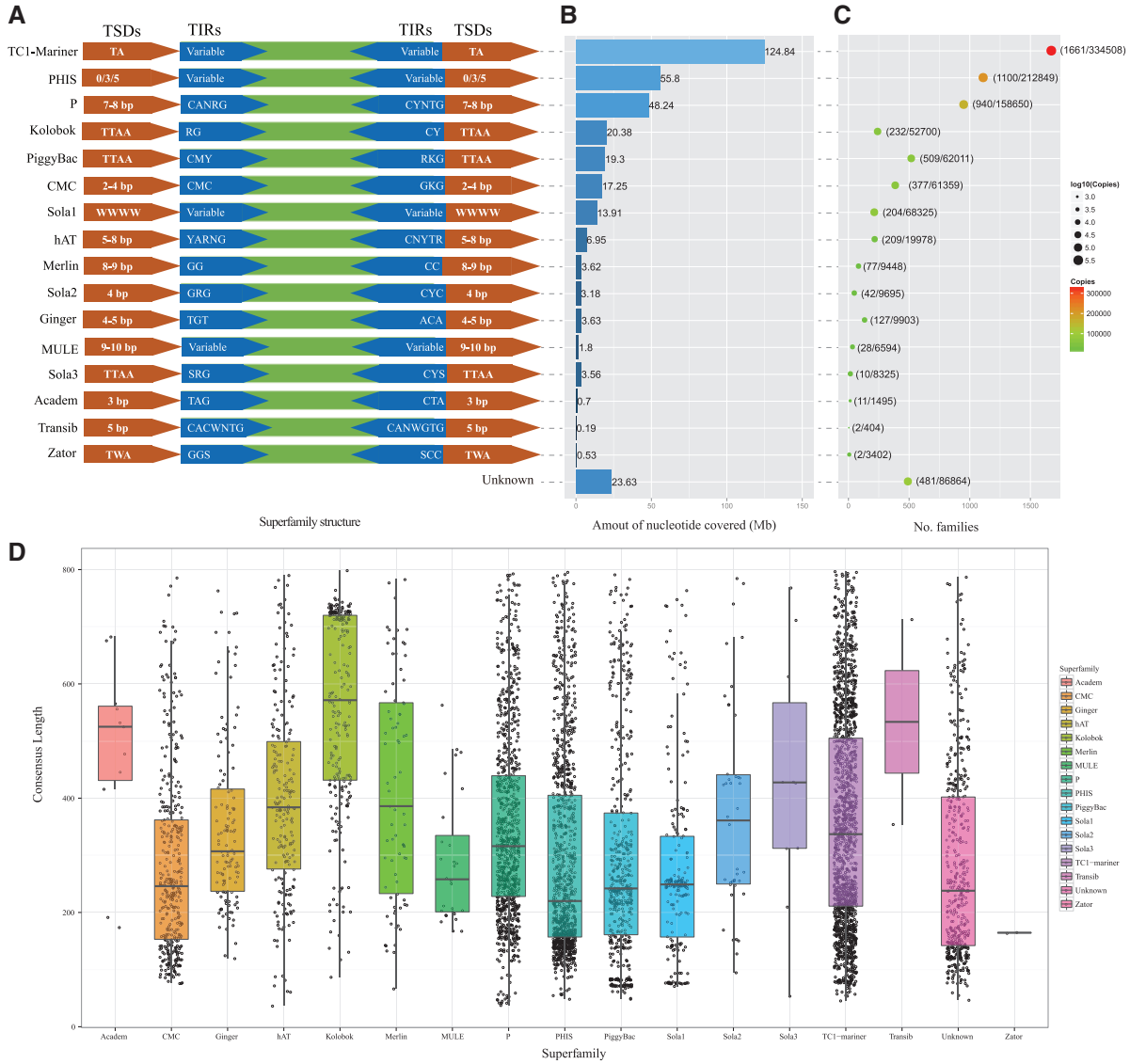

Figure 1. Characteristics of each MITEs superfamily in insect genomes. **(A)** Structure of each superfamily. TSDs sequence and TIRs are shown. **(B)** Amount of nucleotide covered of each superfamily in 98 insect genomes. **(C)** The number of families and copies of each superfamily in the investigated insect genomes. Numbers in parenthesis represents 'families/copies'. **(D)** The distribution of consensus sequence length for each MITE superfamily.

10^{-10}) searches against the Repbase and NCBI non-redundant nucleotide database. Both databases include almost all known MITEs. We found that 5701 (~95%) MITE families did not match to any known TEs in both databases. Therefore, these families were defined as novel MITE families. MITEs like other TEs are huge challenges for host genome sequencing, assembly and annotation due to their repeatability. Thus, the larger number of novel MITE families identified in this study will greatly improve sequencing, assembly and annotation of insect genomes, and facilitate the evolutionary and functional studies of MITEs in the future.

Based on the characteristics of TSDs and TIRs, 5531 MITE families were classified into 16 known superfamilies including *TC1-Mariner*, *PHIS*, *P*, *Kolobok*, *PiggyBac*, *CMC*, *Sola1*, *hAT*, *Merlin*, *Sola2*, *Ginger*, *MULE*, *Sola3*, *Academ*, *Transib* and *Zator*. Meanwhile, 481 families could not be readily assigned to any known DNA transposon superfamilies, and were designated as unknown (Figure 1A). The abundance of different MITE superfamilies in the 98 insect genomes varies markedly. The largest MITE superfamily is the *TC1-Mariner* constituting of 124.84 megabase (Mb) in the insect genomes studied. The smallest superfamily is *Transib* cover 0.19 Mb (Figure 1B). The numbers of families and copies also vary greatly among the different MITE superfamilies. The number of families ranges from 2 to 1,661 and the number of copies ranges from 404 to 334, 508. *TC1-Mariner* is also the largest superfamily

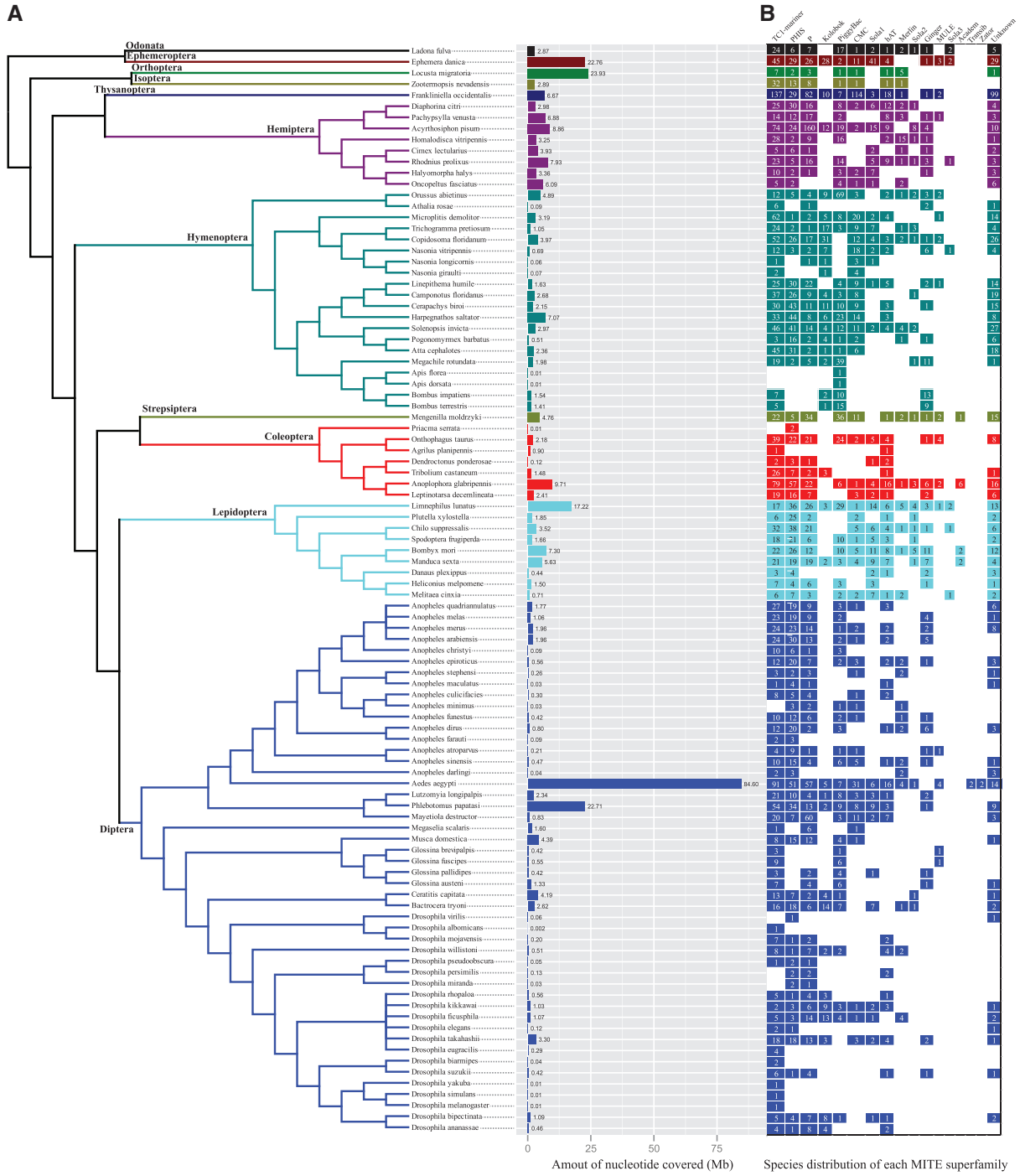

Figure 2. Distribution and abundance of MITEs in 98 insect genomes. **(A)** Amount of nucleotide covered of MITEs in each insect genome. Same color bars represent the same insect order. Numbers represent MITEs abundance (in megabase) in different insect genomes. **(B)** Distribution of MITEs superfamily in each insect genome, color boxes indicated presence; numbers within the color box represent the number of families.

with high numbers of families and copies (Figure 1C). However, the average sequence length of *TC1-Mariner* (~373 bp) superfamily is smaller than *Academ* (~470 bp), *Kolobok* (~386 bp), *Transib* (~482 bp), *Sola3* (~428 bp) and *Merlin* (~383 bp) superfamilies (Figure 1D).

Such abundance of the *TC1-Mariner* superfamily in the insect genomes could be in part explained by its short TSDs because short TSDs likely have much more target sites in host genomes. *TC1-Mariner* transposons are prevalent in eukaryotes, and feature di-nucleotide (5′-TA-3′) TSDs (31). *TC1-Mariner* TSDs are the shortest among known DNA transposon superfamilies (*PiggyBac* is characterized by 5′-TTAA-3′ TSDs, *P* is 7-8 bp TSDs, *Academ* is 3 bp TSDs etc.) (32). TSDs of *Academ* are shorter than *PiggyBac*,

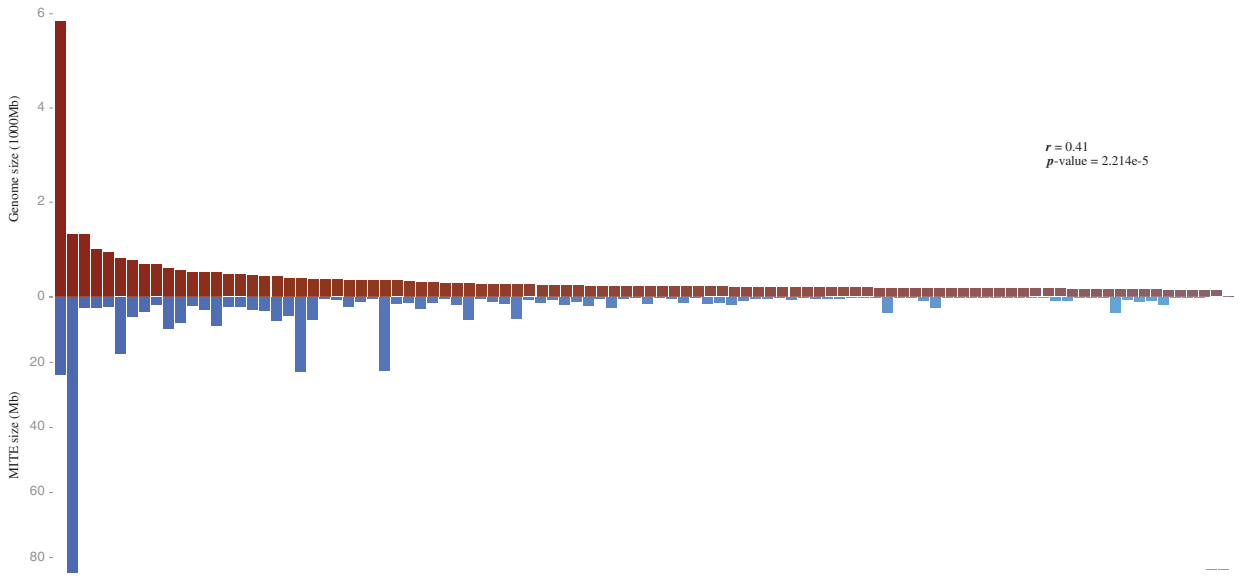

Figure 3. Correlation between the abundance of MITEs and genome size. Histogram above the graph (in red) represents distribution of genome size (unit—1000 megabase). Histogram below the graph (in blue) represents the distribution of MITEs abundance (unit—megabase). Correlation analysis was performed using the R program with the Pearson's method.

Figure 4. The web interface of iMITEdb. The interfaces had browse, search, blast, download, links and contacts.

hAT, *Merlin* etc. However, among the superfamilies, *Academ* has the lowest abundance. Thus, the number of target sites can not completely explain the abundance variation of transposon superfamilies.

The result of correlation analysis revealed that the abundance of each superfamily in insect genomes significantly correlated with the numbers of its families and copies (Supplementary Figure S2A and S2B), and have no significant correlation to its length (Supplementary Figure S2C). In general, the numbers of families and copies of a transposon superfamily are affected by their transposition activities, removal rate or host TE regulation and so on. Whether *TC1-Mariner* in insect genomes has higher transposition activity is to be experimentally verified in the future. If this is case, *TC1-Mariner* could be exploited as a good vector in insect transgenic technology.

MITE abundance in 98 insect genomes

To estimate the abundance of MITEs in each insect genome, the consensus sequence of each MITE family was used as query in a BLASTN search against corresponding genome. The results show that the abundance of MITEs varies greatly among the different genomes (Figure 2A).

For instance, MITEs constitute ~84.60 Mb (occupied ~6.4% of the genome sequence) in the *Aedes aegypti* genome. But *Anopheles darlingii* harbors only ~0.04 Mb (~0.03% of genome) MITE sequences in its genome. Similarly, there are ~2.18 Mb MITE sequences in the *Onthophagus taurus* genome, whereas only ~0.9 Mb in the *Agrilus planipennis* genome. In addition, the numbers of MITE superfamilies and families also vary markedly among the different genomes (Figure 2B). For example, 80 families with 7 known superfamilies were detected in the *Megachile rotundata* genome, whereas only one family was identified in the *Apis florea* genome.

We performed a correlation analysis between the abundance of MITEs in a species and the corresponding genome size and found a very significant correlation between the two characteristics ($r = 0.41$, P-value = 2.214e-5), indicating that the abundance of MITEs influences the insect genome size (Figure 3). This is consistent with the observation in plant genomes (16). When compared with other TEs, MITEs are shorter in a given genome. However, MITEs usually have high copy numbers (12). Thus, they may have an important role in the evolution of the genome size.

Construction of insect MITEs database

We constructed a database, called iMITEdb, using the MITE sequences identified in this study and other available MITE information. The iMITEdb contains the following functions: browse, search, BLAST and download (http://gene.cqu.edu.cn/iMITEdb/) (Figure 4). Each MITE family in the database includes species, superfamily name, family name, TIRs, TSDs, copies, length and consensus sequence. This database allows browsing the information of interesting MITEs based on insect species, superfamily and family. Users can also perform individual search based on a MITE name to obtain information about each MITE family. In BLAST searches, users can enter a sequence in FASTA format or load a DNA sequence containing file to perform BLASTN search against all identified MITEs. All MITE sequences available in the database can also be downloaded. This will greatly facilitate the studies of function and evolution of MITEs in insects in future.

Acknowledgements

We thank all members of Dai's group for their laboratory assistance and useful comments on this article, and thank Dr Cédric Feschotte at Department of Human Genetics, University of Utah, USA, for his helpful discussion during the study.

Funding

This work was supported by the National Natural Science Foundation of China (No. 31401106 to M.J.H., No. 31471197 to Z.Z. and No. 31560308 to Z.H.H.); Fundamental and Advanced Research Project of Chongqing Municipality (No. cstc2016jcyjA0258 to M.J.H.); the Hi-Tech Research and Development (863) Program of China (No. 2013AA102507 to F.Y.D.); Fundamental Research Funds for the Central Universities (XDJK2016C009 and SWU115035 to M.J.H.).

Supplementary data

Supplementary data are available at *Database* Online.

Conflict of interest: None declared.

References

1. Wessler,S.R. and Varagona,M.J. (1985) Molecular basis of mutations at the waxy locus of maize: correlation with the fine structure genetic map. *Proc. Natl. Acad. Sci. USA*, 82, 4177–4181.
2. Bureau,T.E. and Wessler,S.R. (1992) Tourist: a large family of small inverted repeat elements frequently associated with maize genes. *Plant Cell*, 4, 1283–1249.
3. Tu,Z. (2001) Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, *Anopheles gambiae*. *Proc. Natl. Acad. Sci. USA*, 98, 1699–1704.
4. Oki,N., Yano,K., Okumoto,Y. *et al.* (2008) A genome-wide view of miniature inverted-repeat transposable elements (MITEs) in rice, *Oryza sativa ssp. japonica*. *Genes Genet. Syst.*, 83, 321–329.
5. Han,M.J., Shen,Y.H., Gao,Y.H. *et al.* (2010) Burst expansion, distribution and diversification of MITEs in the silkworm genome. *BMC Genomics*, 11, 520.
6. Feschotte,C. and Mouchès,C. (2000) Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a pogo-like DNA transposon. *Mol. Biol. Evol*, 17, 730–737.
7. Feschotte,C., Swamy,L., and Wessler,S.R. (2003) Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (MITEs). *Genetics*, 163, 747–758.
8. Feschotte,C., Jiang,N., and Wessler,S.R. (2002) Plant transposable elements: where genetics meets genomics. *Nat. Rev. Genet.*, 3, 329–341.
9. Yang,G., Nagel,D.H., Feschotte,C. *et al.* (2009) Tuned for transposition: molecular determinants underlying the hyperactivity of a Stowaway MITE. *Science*, 325, 1391–1394.
10. Bureau,T.E. and Wessler,S.R. (1994) Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell*, 6, 907–916.
11. Smit,A.F. and Riggs,A.D. (1996) Tiggers and DNA transposon fossils in the human genome. *Proc. Natl. Acad. Sci. USA*, 93, 1443–1448.
12. Jiang,N., Bao,Z., Zhang,X. *et al.* (2003) An active DNA transposon family in rice. *Nature*, 421, 163–167.
13. Wessler,S.R., Bureau,T.E., and White,S.E. (1995) LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Curr. Opin. Genet. Dev*, 5, 814–821.
14. Naito,K., Zhang,F., Tsukiyama,T. *et al.* (2009) Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature*, 461, 1130–1134.
15. Chen,J., Lu,C., Zhang,Y. *et al.* (2012) Miniature inverted-221 repeat transposable elements (MITEs) in rice were originated and amplified predominantly after the divergence of Oryza and

Brachypodium and contributed considerable diversity to the species. *Mob. Genet. Elements*, 2, 127–132.
16. Chen,J., Hu,Q., Zhang,Y. *et al.* (2014) P-MITE: a database for plant miniature inverted-repeat transposable elements. *Nucleic Acids Res.*, 42, D1176–D1181.
17. Yang,G., Lee,Y.H., Jiang,Y. *et al.* (2005) A two-edged role for the transposable element *Kiddo* in the rice ubiquitin2 promoter. *Plant Cell*, 17, 1559–1568.
18. Hollister,J.D. and Gaut,B.S. (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res.*, 19, 1419–1428.
19. Kuang,H., Padmanabhan,C., Li,F. *et al.* (2009) Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: new functional implications for MITEs. *Genome Res.*, 19, 42–56.
20. Lu,C., Chen,J., Zhang,Y. *et al.* (2012) 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.*, 29, 1005–1017.
21. Amundsen,K., Rotter,D., Li,H. *et al.* (2011) Miniature inverted-repeat transposable element identification and genetic marker development in Agrostis. *Crop Sci.*, 51, 854–861.
22. Hancock,C.N., Zhang,F., Floyd,K. *et al.* (2011) The rice miniature inverted repeat transposable element *mPing* is an effective insertional mutagen in soybean. *Plant Physiol.*, 157, 552–562.
23. Yaakov,B., Ceylan,E., Domb,K. *et al.* (2012) Marker utility of miniature inverted-repeat transposable elements for wheat biodiversity and evolution. *Theor. Appl. Genet.*, 124, 1365–1373.
24. Fattash,I., Bhardwaj,P., Hui,C. *et al.* (2013) A rice stowaway MITE for gene transfer in yeast. *PLoS One*, 8, e64135.
25. Yang,G. and Hall,T.C. (2003) MAK, a computational tool kit for automated MITE analysis. *Nucleic Acids Res.*, 31, 3659–3665.
26. Chen,Y., Zhou,F., Li,G. *et al.* (2009) MUST: a system for identification of miniature inverted-repeat transposable elements and applications to *Anabaena variabilis* and *Haloquadratum walsbyi*. *Gene*, 436, 1–7.
27. Han,Y., and Wessler,S.R. (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.*, 38, e199.
28. Yang,G. (2013) MITE Digger, an efficient and accurate algorithm for genome wide discovery of miniature inverted repeat transposable elements. *BMC Bioinformatics*, 14, 186.
29. Nene,V., Wortman,J.R., Lawson,D. *et al.* (2007) Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science*, 316, 1718–1723.
30. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32, 1792–1797.
31. Shao,H., and Tu,Z. (2001) Expanding the diversity of the IS630-Tc1-mariner superfamily: discovery of a unique DD37E transposon and reclassification of the DD37D and DD39D transposons. *Genetics*, 159, 1103–1115.
32. Yuan,Y.W. and Wessler,S.R. (2011) The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc. Natl. Acad. Sci. USA*, 108, 7884–7889.