D1176–D1181 *Nucleic Acids Research, 2014, Vol. 42, Database issue* *Published online 29 October 2013*
doi:10.1093/nar/gkt1000

P-MITE: a database for plant miniature inverted-repeat transposable elements

Jiongjiong Chen, Qun Hu, Yu Zhang, Chen Lu and Hanhui Kuang*

Department of Vegetable Crops, Key Laboratory of Horticulture Biology, Ministry of Education, College of Horticulture and Forestry Sciences, Huazhong Agricultural University, Wuhan, 430070, P. R. China

Received June 2, 2013; Revised September 13, 2013; Accepted October 4, 2013

ABSTRACT

Miniature inverted-repeat transposable elements (MITEs) are prevalent in eukaryotic species including plants. MITE families vary dramatically and usually cannot be identified based on homology. In this study, we *de novo* identified MITEs from 41 plant species, using computer programs MITE Digger, MITE-Hunter and/or Repetitive Sequence with Precise Boundaries (RSPB). MITEs were found in all, but one (*Cyanidioschyzon merolae*), species. Combined with the MITEs identified previously from the rice genome, >2.3 million sequences from 3527 MITE families were obtained from 41 plant species. In general, higher plants contain more MITEs than lower plants, with a few exceptions such as papaya, with only 538 elements. The largest number of MITEs is found in apple, with 237 302 MITE sequences. The number of MITE sequences in a genome is significantly correlated with genome size. A series of databases (plant MITE databases, P-MITE), available online at http://pmite.hzau.edu.cn/django/mite/, was constructed to host all MITE sequences from the 41 plant genomes. The databases are available for sequence similarity searches (BLASTN), and MITE sequences can be downloaded by family or by genome. The databases can be used to study the origin and amplification of MITEs, MITE-derived small RNAs and roles of MITEs on gene and genome evolution.

INTRODUCTION

Miniature inverted-repeat transposable elements (MITEs) are prevalent in eukaryotic genomes, and are believed to be deletion derivatives of DNA transposons (1,2). Like autonomous DNA transposons, MITEs usually have terminal inverted repeats (TIR), flanked by short direct repeats [also called target site duplication (TSD)]. Compared with autonomous DNA transposons, MITEs are often short (<800 bp) and do not encode transposases.

MITEs are often located in gene-rich euchromatic regions and are associated with genes (3,4). Several pieces of evidence suggest that MITEs may affect the expression of nearby genes. MITE *Kiddo* in rice was shown to upregulate the expression of *Ubiquitin2* when inserted in its promoter region (5). However, in other cases, MITE insertions downregulate the expression of nearby genes (6,7). Such downregulation is most likely through small RNAs derived from MITE sequences (6,8). MITE transpositions generate much genetic diversity for a species (9–11). Considering the effects of MITEs on gene expression and variation of MITE insertions in different genotypes, MITEs may contribute to considerable phenotypic diversity as well (12).

The first MITE families were discovered through sequence analysis (i.e. identification of TIR and TSD sequences) of insertions of 100–600 bp (13,14). Recently, computer programs were developed to systematically identify MITEs from a database such as genome sequences (6,15–19). Among them, the most successful ones are MITE Digger, MITE-Hunter and RSPB, which identified the vast majority of MITEs in the sequenced genome of rice (6,18,19). The recently reported program MITE Digger is most efficient for *de novo* MITE identification, particularly in large genomes (19). RSPB is better at identifying MITE families with atypical structures such as MITEs with no TSD or short/diverse TIR sequences. Unfortunately, RSPB requires high computer capacity not found in most laboratories. We predicted that combining MITE Digger, MITE-Hunter and RSPB would allow the detection of a vast majority of, if not all, MITE families in a genome, with no prior information required. With the availability of the three MITE detecting programs and the genome sequences of many plant species, MITEs in several genomes can be readily identified and compared

*To whom correspondence should be addressed. Tel: +86 27 87280752; Fax: +86 27 87282010; Email: kuangfile@gmail.com

The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

to further our understanding of MITE origin and evolution.

MITEs, as repetitive sequences, were included in other databases such as the The Institute for Genomic Research (TIGR) Plant Repeat Databases and Repbase (20,21). However, MITEs vary dramatically and usually cannot be identified through homology search between distantly related species, and consequently, only a small proportion of MITE families have been identified and included in these databases. In this study, MITEs were *de novo* identified from 41 plant species using computer programs MITE Digger, MITE-Hunter and/or RSPB. Each MITE family was annotated manually. All verified MITE families were stored in a database, P-MITE (for plant MITE). BLASTN search function was appended into the database. MITE sequences from each genome were downloadable. P-MITE will be helpful for the annotation of genes and genomic sequences. It can also be used to study the origin and amplification of MITEs, the comparative analysis between different species, the MITE-derived small RNAs and the roles of MITEs on gene and genome evolution, etc.

MATERIALS AND METHODS

Plant genomes used in this study

Forty-one sequenced and published genomes of plant species, including six lower plant species, were included in this study for MITE identification. The information of the 41 species and the Web sites for their genome sequences are listed in Supplementary Table S1. The MITEs from rice were identified and annotated in a previous study (6).

De novo identification of MITEs using MITE Digger, MITE-Hunter and RSPB

MITEs from 41 genomes were *de novo* identified using program MITE Digger, MITE-hunter and/or RSPB (6,18,19). First, program MITE-Hunter was used to run the sequences of each genome. The resulting groups of potential MITEs were manually checked for TSD and TIR sequences. Groups with no precise boundaries (terminals) or no TIR sequences were not considered as MITEs. The confirmed MITEs from MITE-Hunter were put into a database (MITE-Hunter database). To save running time, program RSPB was slightly modified so that the confirmed MITE sequences in the 'MITE-Hunter database' were skipped by RSPB. New groups of repetitive sequences with precise boundaries were reported and checked manually for TSDs and TIRs (Supplementary Figure S1). No TSD and TIR information is required to run RSPB, which identifies repetitive sequences with precise boundaries. In subsequent manual annotation, only repetitive sequences <800 bp and TSD/TIR features similar to known MITE superfamilies were maintained. Five species with large genomes or too many short contigs were not successful using RSPB. MITE Digger, released recently, was also used to run some genomes, including genomes >800 Mb. The statistics of MITE families identified in this study is shown in Supplementary Table S2. The number of MITE families that were detected by RSPB, but not by MITE Hunter, is shown in Supplementary Table S3.

Classification of MITE superfamily and family

A Perl script was written to cluster MITEs identified above into a family if they had significant sequence similarity (BLASTN $e < 10^{-10}$) (6). MITE families were assigned into superfamilies based on their TIR and TSD sequences. Each MITE family in a genome was named as code_Abc#, where Ab is the first two letters from its genus name, c the first letter from its species name and # a consecutive number. Different superfamilies are represented by different codes, with DTT for *Tc1/Mariner*, DTM for *Mutator*, DTA for *h*AT, DTC for *CACTA*, DTH for *PIF/Harbinger*, DTP for *P*, DTN for *Novosib* and DTx for unknown (21–23). MITEs with ambiguous TSD and/or TIR features were annotated as unknown superfamily (DTx). MITE families preferentially inserted into simple tandem repeats (microsatellites) were considered as an independent group, *MiM* (MITEs inserted in microsatellite). A 'representative' element was chosen for each family, and the representative elements should have good TIR and perfect TSD sequences if possible. A MITE sequence was considered as a full-length element when its terminals were no more than 3 bp shorter than the representative sequence. To identify all MITE elements, including diverse and/or partial ones, in a genome, a library of all representative elements from each family was used as query sequences to search the entire genome sequence using RepeatMasker v3.2.9 (http://www.repeatmasker.org/).

RESULTS AND DISCUSSION

De novo identification of MITEs in 41 plant genomes

Program MITE-Hunter was applied to 41 plant genomes for genome-wide *de novo* identification of MITEs. RSPB was also used to run all but five genomes that are either >800 Mb or with too many contigs. MITE Digger was used to search some genomes, including four skipped by RSPB. The MITE sequences obtained from this study were used to execute a BLASTN search of the Repbase, the database most frequently used for repetitive sequences (21). More than 70% of MITE families identified from this study were not included in Repbase ($< 10^{-10}$), MITE-Hunter, but not RSPB, due to too large genome. A total of 252 MITE families were obtained from maize, which include 97 novel families not covered by maize TE database. However, 61 MITE families listed in maize TE database were not identified by either MITE Digger or MITE-Hunter. The computing process of RSPB needs to be mended before it can be applied to large genomes, such as maize, to identify more novel MITE families.

The majority of MITEs were classified into five superfamilies, including *Tc1/Mariner*, *PIF/Harbinger*, *C ACTA*, *h*AT and *Mutator*. Two superfamilies, *P* and *Novosib*, were detected in the genomes of lower plants, although they do not have *Tc1/Mariner*, *CACTA* and *Mutator*. Sixteen MITE families were unclassified owing to ambiguous TSD and/or TIR features. *MiM* is the least frequent in plant genomes (Supplementary Table S2). The *MiM* group is present in only 10 of the 41 genomes, with 41 893 elements from 33 families. The strawberry genome contains 14 *MiM* families, whereas the others have no more than four *MiM* families. Most elements of these *MiM* families, including the *Micron* family in rice (24), were inserted in $(TA)_n$ repeats, with only a few exceptions, in which they were inserted into $(CA)_n/(GT)_n$ repeats. Elements from the *MiM* group have poor TIR sequences, and no conserved nucleotides were found in their terminals among different families. It remains unclear whether different *MiM* families belong to the same superfamily, i.e. activated by the same type of transposase. In contrast to the scarce *MiM* group, the *Mutator* superfamily has 852 390 elements in the 41 genomes included in this study, with an average of >20 790 elements per genome.

MITEs with significant nucleotide identities (BLASTN $e < 10^{-10}$) were grouped into a family. The largest MITE family is the *DTM_Mad25* from the apple genome, with 18 904 elements. The smallest MITE families, *DTT_Sob24* and *DTH_Sob33* from the *Sorghum* genome, have only one element.

The number of MITEs varies dramatically in different species. In general, the genomes of lower plants have relatively few MITEs (Table 1). No MITEs were detected in the genome of *Cyanidioschyzon merolae* using either MITE-Hunter or RSPB, and the genome of *Selaginella moellendorffii* harbors only 73 MITE elements. The number of MITEs also varies considerably among the genomes of higher plants. For example, only one MITE family with 538 elements was detected in the papaya genome, whereas 237 302 elements from 180 MITE families are present in the apple genome. Large variations

Table 1. MITE in 41 plant genomes

Species	Family	Genome size (Mb)	MITE			
			Family number	Element number	Total length (Mb)	Percentage in genome
Phoenix dactylifera	Arecaceae	381.56	33	39 990	8.22	2.15
Arabidopsis thaliana	Brassicaceae	119.67	43	3245	0.85	0.71
Thellungiella parvula	Brassicaceae	123.6	7	1161	0.32	0.26
Arabidopsis lyrata	Brassicaceae	206.67	121	18 039	4.64	2.24
Thellungiella salsuginea	Brassicaceae	208.87	54	5133	1.27	0.61
Brassica rapa	Brassicaceae	283.84	174	45 821	11.49	4.05
Carica papaya	Caricaceae	342.68	1	538	0.21	0.06
Chlamydomonas reinhardtii	Chlamydomonadaceae	111.1	20	3508	0.99	0.89
Chlorella variabilis	Chlorellaceae	46.16	2	83	0.04	0.08
Cucumis sativus	Cucurbitaceae	203.06	7	10 810	2.02	1.00
Citrullus lanatus	Cucurbitaceae	353.47	35	94 314	19.55	5.53
Cucumis melo	Cucurbitaceae	431.04	10	12 991	2.79	0.65
Cyanidioschyzon merolae	Cyanidiaceae	16.54	0	0	0.00	0.00
Jatropha curcas	Euphorbiaceae	297.67	17	18 975	4.81	1.61
Ricinus communis	Euphorbiaceae	350.63	33	13 205	3.24	0.93
Manihot esculenta	Euphorbiaceae	532.53	21	30 934	8.94	1.68
Medicago truncatula	Fabaceae	307.48	288	132 834	25.24	8.21
Lotus japonicus	Fabaceae	316.89	172	71 811	14.16	4.47
Cajanus cajan	Fabaceae	605.78	92	135 581	31.06	5.13
Cannabis sativa	Fabaceae	786.64	53	110 123	24.06	3.06
Glycine max	Fabaceae	973.34	126	169 379	27.69	2.84
Physcomitrella patens	Funariaceae	479.99	4	3718	0.58	0.12
Linum usitatissimum	Linaceae	318.25	28	14 409	3.51	1.10
Theobroma cacao	Malvaceae	327.35	13	10 364	3.45	1.06
Musa acuminate	Musaceae	472.96	9	15 835	2.22	0.47
Coccomyxa subellipsoidea	Palmellaceae	48.95	4	187	0.04	0.09
Brachypodium distachyon	Poaceae	271.92	222	83 272	12.86	4.73
Oryza sativa[a]	Poaceae	373.25	339	179 415	37.27	9.98
Setaria italica	Poaceae	405.78	178	69 264	15.60	3.85
Sorghum bicolor	Poaceae	738.58	275	112 307	29.63	4.01
Zea mays	Poaceae	2058.58	252	192 529	40.36	1.96
Fragaria vesca	Rosaceae	206.89	162	34 880	8.97	4.33
Malus domestica	Rosaceae	881.28	180	237 302	44.63	5.06
Prunus persica	Rosaceae	227.25	99	39 110	8.84	3.89
Citrus sinensis	Rutaceae	327.94	106	46 032	11.35	3.46
Populus trichocarpa	Salicaceae	417.14	22	35 081	7.49	1.80
Selaginella moellendorffii	Selaginellaceae	212.76	1	73	0.01	0.01
Solanum lycopersicum	Solanaceae	781.67	104	107 087	26.89	3.44
Solanum tuberosum	Solanaceae	797.83	171	170 392	38.65	4.84
Vitis vinifera	Vitaceae	486.19	35	61 065	14.69	3.02
Volvox carteri	Volvocaceae	131.16	14	2104	0.62	0.47

[a] The MITE sequences from rice were retrieved from Lu *et al.* (25).

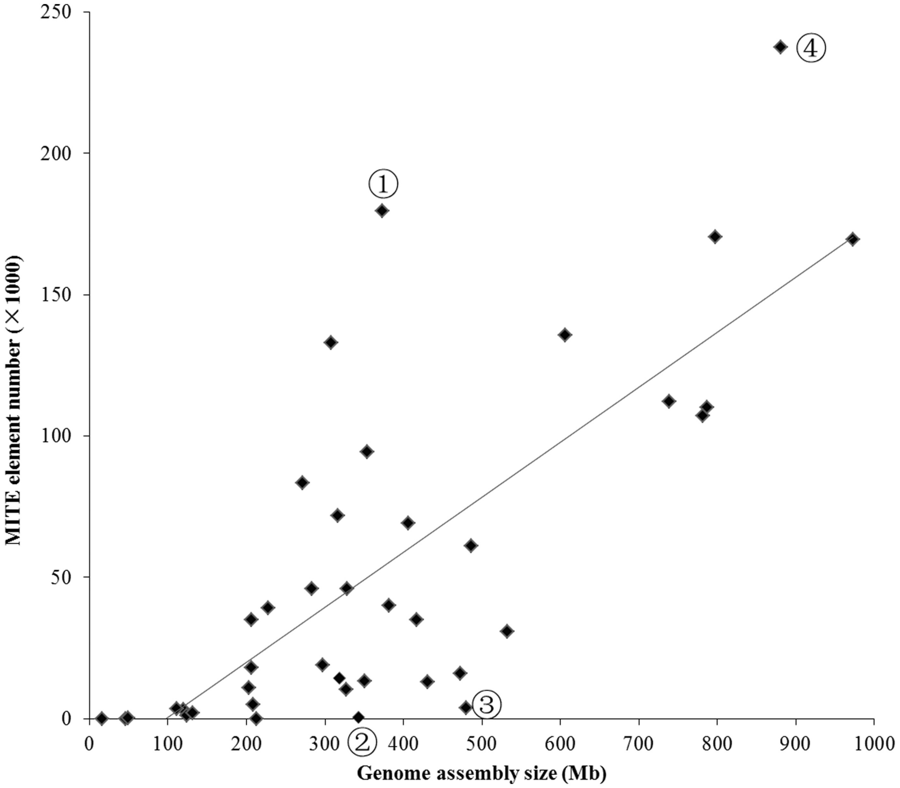

Figure 1. Strong correlation between the number of MITEs and genome assembly size. Genomes with disproportionately low copy (② papaya and ③ *Physcomitrella patens*) and high copy (① rice and ④ apple) of MITEs are indicated.

in total number of MITE elements also occur between closely related species. For example, the *Arabidopsis thaliana* genome has only 3245 MITE elements, whereas its close relative, *Arabidopsis lyrata,* contains 18 039 MITE-related sequences. Similarly, the number of MITEs in the genome of watermelon (with 94 314 MITE elements) is seven times as much as in the genome of melon (with 12 991 MITE elements).

The number of MITEs in a genome is significantly correlated with its genome assembly size ($r = 0.72$, $P < 0.01$; Table 1; Figure 1). A similar correlation coefficient ($r = 0.68$, $P < 0.01$) was obtained when the six lower plants were excluded from the analysis. Nevertheless, several striking exceptions were observed. For example, the rice genome is only 373 Mb but has the third largest number (179 415) of MITEs among all species studied, whereas papaya with genome size (342 Mb) similar to that of rice, has only 538 elements of one MITE family (Table 1).

The construction and the use of plant MITE database, P-MITE

A total of 2.3 million sequences of 3527 MITE families were obtained from 41 (including the rice genome) plant genomes. A series of databases containing MITEs from the 41 plant genomes was constructed. Elements from each of the 3527 MITE families were checked and annotated manually, and one element with better TSD and/or TIR features was chosen as a representative of the family. A database containing all representative elements was constructed, which can be used to study the structure of MITEs, such as their TSD and TIR features.

The aforementioned databases are collectively named as P-MITE (for plant MITE), and can be found in http://pmite.hzau.edu.cn/django/mite. The database is searchable using BLASTN algorithm. MITE sequences and representative elements can be downloaded by family or by genome.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online, including [26–66].

ACKNOWLEDGEMENTS

The authors are grateful to all the members in the Plant Genomics Lab, College of Horticulture and Forestry, Huazhong Agricultural University, for their manual annotation of MITE families.

FUNDING

This work was supported by the ‘973’ National Key Basic Research Program [2009CB119000]; The National Natural Science Foundation of China (NSFC) [30921002]; and the Fundamental Research Funds for the Central Universities [2012ZYTS035]. Funding for open access charge: The Fundamental Research Funds for the Central Universities [2012ZYTS035].

Conflict of interest statement. None declared.

REFERENCES

1. Feschotte,C. and Mouches,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.
2. 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.
3. Han,Y., Qin,S. and Wessler,S.R. (2013) Comparison of class 2 transposable elements at superfamily resolution reveals conserved and distinct features in cereal grass genomes. *BMC Genomics*, **14**, 71.
4. Tu,Z. (1997) Three novel families of miniature inverted-repeat transposable elements are associated with genes of the yellow fever mosquito, Aedes aegypti. *Proc. Natl Acad. Sci. USA*, **94**, 7475–7480.
5. Yang,G., Lee,Y.H., Jiang,Y., Shi,X., Kertbundit,S. and Hall,T.C. (2005) A two-edged role for the transposable element Kiddo in the rice ubiquitin2 promoter. *Plant Cell*, **17**, 1559–1568.
6. Lu,C., Chen,J.J., Zhang,Y., Hu,Q., Su,W.Q. and Kuang,H.H. (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.
7. 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.
8. Kuang,H., Padmanabhan,C., Li,F., Kamei,A., Bhaskar,P.B., Ouyang,S., Jiang,J., Buell,C.R. and Baker,B. (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.
9. Casa,A.M., Brouwer,C., Nagel,A., Wang,L., Zhang,Q., Kresovich,S. and Wessler,S.R. (2000) The MITE family heartbreaker (Hbr): molecular markers in maize. *Proc. Natl Acad. Sci. USA*, **97**, 10083–10089.
10. Shirasawa,K., Hirakawa,H., Tabata,S., Hasegawa,M., Kiyoshima,H., Suzuki,S., Sasamoto,S., Watanabe,A., Fujishiro,T. and Isobe,S. (2012) Characterization of active miniature inverted-repeat transposable elements in the peanut genome. *Theor. Appl. Genet.*, **124**, 1429–1438.
11. Yaakov,B., Ceylan,E., Domb,K. and Kashkush,K. (2012) Marker utility of miniature inverted-repeat transposable elements for wheat biodiversity and evolution. *Theor. Appl. Genet.*, **124**, 1365–1373.
12. Chen,J., Lu,C., Zhang,Y. and Kuang,H. (2012) Miniature inverted-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.
13. 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**, 1283–1294.
14. 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.
15. 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.
16. Santiago,N., Herraiz,C., Goni,J.R., Messeguer,X. and Casacuberta,J.M. (2002) Genome-wide analysis of the Emigrant family of MITEs of Arabidopsis thaliana. *Mol. Biol. Evol.*, **19**, 2285–2293.
17. Chen,Y., Zhou,F., Li,G. and Xu,Y. (2009) MUST: a system for identification of miniature inverted-repeat transposable elements and applications to Anabaena variabilis and Haloquadratum walsbyi. *Gene*, **436**, 1–7.
18. 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.
19. Yang,G. (2013) MITE Digger, an efficient and accurate algorithm for genome wide discovery of miniature inverted repeat transposable elements. *BMC Bioinformatics*, **14**, 186.
20. Ouyang,S. and Buell,C.R. (2004) The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res.*, **32**, D360–D363.
21. Jurka,J., Kapitonov,V.V., Pavlicek,A., Klonowski,P., Kohany,O. and Walichiewicz,J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, **110**, 462–467.
22. Kapitonov,V.V. and Jurka,J. (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genet.*, **9**, 411–412, author reply 414.
23. Wicker,T., Sabot,F., Hua-Van,A., Bennetzen,J.L., Capy,P., Chalhoub,B., Flavell,A., Leroy,P., Morgante,M., Panaud,O. *et al.* (2007) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.*, **8**, 973–982.
24. Akagi,H., Yokozeki,Y., Inagaki,A., Mori,K. and Fujimura,T. (2001) Micron, a microsatellite-targeting transposable element in the rice genome. *Mol. Genet. Genomics*, **266**, 471–480.
25. Lu,C., Chen,J., Zhang,Y., Hu,Q., Su,W. and Kuang,H. (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.
26. Al-Dous,E.K., George,B., Al-Mahmoud,M.E., Al-Jaber,M.Y., Wang,H., Salameh,Y.M., Al-Azwani,E.K., Chaluvadi,S., Pontaroli,A.C., DeBarry,J. *et al.* (2011) De novo genome sequencing and comparative genomics of date palm (Phoenix dactylifera). *Nat. Biotechnol.*, **29**, 521–527.
27. Hu,T.T., Pattyn,P., Bakker,E.G., Cao,J., Cheng,J.F., Clark,R.M., Fahlgren,N., Fawcett,J.A., Grimwood,J., Gundlach,H. *et al.* (2011) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.*, **43**, 476–481.
28. Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796–815.
29. Wang,X., Wang,H., Wang,J., Sun,R., Wu,J., Liu,S., Bai,Y., Mun,J.H., Bancroft,I., Cheng,F. *et al.* (2011) The genome of the mesopolyploid crop species Brassica rapa. *Nat. Genet.*, **43**, 1035–1039.
30. Dassanayake,M., Oh,D.H., Haas,J.S., Hernandez,A., Hong,H., Ali,S., Yun,D.J., Bressan,R.A., Zhu,J.K., Bohnert,H.J. *et al.* (2011) The genome of the extremophile crucifer Thellungiella parvula. *Nat. Genet.*, **43**, 913–918.
31. Wu,H.J., Zhang,Z.H., Wang,J.Y., Oh,D.H., Dassanayake,M., Liu,B.H., Huang,Q.F., Sun,H.X., Xia,R., Wu,Y.R. *et al.* (2012) Insights into salt tolerance from the genome of Thellungiella salsuginea. *Proc. Natl Acad. Sci. USA*, **109**, 12219–12224.
32. Ming,R., Hou,S., Feng,Y., Yu,Q., Dionne-Laporte,A., Saw,J.H., Senin,P., Wang,W., Ly,B.V., Lewis,K.L. *et al.* (2008) The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature*, **452**, 991–996.
33. Merchant,S.S., Prochnik,S.E., Vallon,O., Harris,E.H., Karpowicz,S.J., Witman,G.B., Terry,A., Salamov,A., Fritz-Laylin,L.K., Marechal-Drouard,L. *et al.* (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science*, **318**, 245–250.
34. Blanc,G., Duncan,G., Agarkova,I., Borodovsky,M., Gurnon,J., Kuo,A., Lindquist,E., Lucas,S., Pangilinan,J., Polle,J. *et al.* (2010) The Chlorella variabilis NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. *Plant Cell*, **22**, 2943–2955.
35. Guo,S., Zhang,J., Sun,H., Salse,J., Lucas,W.J., Zhang,H., Zheng,Y., Mao,L., Ren,Y., Wang,Z. *et al.* (2013) The draft genome of watermelon (Citrullus lanatus) and resequencing of 20 diverse accessions. *Nat. Genet.*, **45**, 51–58.

36. Garcia-Mas,J., Benjak,A., Sanseverino,W., Bourgeois,M., Mir,G., Gonzalez,V.M., Henaff,E., Camara,F., Cozzuto,L., Lowy,E. *et al.* (2012) The genome of melon (Cucumis melo L.*). Proc. Natl Acad. Sci. USA*, **109**, 11872–11877.
37. Huang,S., Li,R., Zhang,Z., Li,L., Gu,X., Fan,W., Lucas,W.J., Wang,X., Xie,B., Ni,P. *et al.* (2009) The genome of the cucumber, Cucumis sativus L. *Nat. Genet.*, **41**, 1275–1281.
38. Matsuzaki,M., Misumi,O., Shin,I.T., Maruyama,S., Takahara,M., Miyagishima,S.Y., Mori,T., Nishida,K., Yagisawa,F., Nishida,K. *et al.* (2004) Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D. *Nature*, **428**, 653–657.
39. Sato,S., Hirakawa,H., Isobe,S., Fukai,E., Watanabe,A., Kato,M., Kawashima,K., Minami,C., Muraki,A., Nakazaki,N. *et al.* (2011) Sequence analysis of the genome of an oil-bearing tree, Jatropha curcas L. *DNA Res.*, **18**, 65–76.
40. Prochnik,S., Marri,P.R., Desany,B., Rabinowicz,P.D., Kodira,C., Mohiuddin,M., Rodriguez,F., Fauquet,C., Tohme,J., Harkins,T. *et al.* (2012) The cassava genome: current progress, future directions. *Trop. Plant Biol.*, **5**, 88–94.
41. Chan,A.P., Crabtree,J., Zhao,Q., Lorenzi,H., Orvis,J., Puiu,D., Melake-Berhan,A., Jones,K.M., Redman,J., Chen,G. *et al.* (2010) Draft genome sequence of the oilseed species Ricinus communis. *Nat. Biotechnol.*, **28**, 951–956.
42. Varshney,R.K., Chen,W., Li,Y., Bharti,A.K., Saxena,R.K., Schlueter,J.A., Donoghue,M.T., Azam,S., Fan,G., Whaley,A.M. *et al.* (2012) Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.*, **30**, 83–89.
43. van Bakel,H., Stout,J.M., Cote,A.G., Tallon,C.M., Sharpe,A.G., Hughes,T.R. and Page,J.E. (2011) The draft genome and transcriptome of Cannabis sativa. *Genome Biol.*, **12**, R102.
44. Schmutz,J., Cannon,S.B., Schlueter,J., Ma,J., Mitros,T., Nelson,W., Hyten,D.L., Song,Q., Thelen,J.J., Cheng,J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.
45. Sato,S., Nakamura,Y., Kaneko,T., Asamizu,E., Kato,T., Nakao,M., Sasamoto,S., Watanabe,A., Ono,A., Kawashima,K. *et al.* (2008) Genome structure of the legume, Lotus japonicus. *DNA Res.*, **15**, 227–239.
46. Young,N.D., Debelle,F., Oldroyd,G.E.D., Geurts,R., Cannon,S.B., Udvardi,M.K., Benedito,V.A., Mayer,K.F.X., Gouzy,J., Schoof,H. *et al.* (2011) The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature*, **480**, 520–524.
47. Rensing,S.A., Lang,D., Zimmer,A.D., Terry,A., Salamov,A., Shapiro,H., Nishiyama,T., Perroud,P.F., Lindquist,E.A., Kamisugi,Y. *et al.* (2008) The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science*, **319**, 64–69.
48. Wang,Z., Hobson,N., Galindo,L., Zhu,S., Shi,D., McDill,J., Yang,L., Hawkins,S., Neutelings,G., Datla,R. *et al.* (2012) The genome of flax (Linum usitatissimum) assembled de novo from short shotgun sequence reads. *Plant J.*, **72**, 461–473.
49. Argout,X., Salse,J., Aury,J.M., Guiltinan,M.J., Droc,G., Gouzy,J., Allegre,M., Chaparro,C., Legavre,T., Maximova,S.N. *et al.* (2011) The genome of Theobroma cacao. *Nat. Genet.*, **43**, 101–108.
50. D'Hont,A., Denoeud,F., Aury,J.M., Baurens,F.C., Carreel,F., Garsmeur,O., Noel,B., Bocs,S., Droc,G., Rouard,M. *et al.* (2012) The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature*, **488**, 213–217.
51. Blanc,G., Agarkova,I., Grimwood,J., Kuo,A., Brueggeman,A., Dunigan,D.D., Gurnon,J., Ladunga,I., Lindquist,E., Lucas,S. *et al.* (2012) The genome of the polar eukaryotic microalga Coccomyxa subellipsoidea reveals traits of cold adaptation. *Genome Biol.*, **13**, R39.
52. International Brachypodium Initiative. (2010) Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature*, **463**, 763–768.
53. International Rice Genome Sequencing Project. (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
54. Zhang,G., Liu,X., Quan,Z., Cheng,S., Xu,X., Pan,S., Xie,M., Zeng,P., Yue,Z., Wang,W. *et al.* (2012) Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential. *Nat. Biotechnol.*, **30**, 549–554.
55. Paterson,A.H., Bowers,J.E., Bruggmann,R., Dubchak,I., Grimwood,J., Gundlach,H., Haberer,G., Hellsten,U., Mitros,T., Poliakov,A. *et al.* (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551–556.
56. Schnable,P.S., Ware,D., Fulton,R.S., Stein,J.C., Wei,F., Pasternak,S., Liang,C., Zhang,J., Fulton,L., Graves,T.A. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.
57. Velasco,R., Zharkikh,A., Affourtit,J., Dhingra,A., Cestaro,A., Kalyanaraman,A., Fontana,P., Bhatnagar,S.K., Troggio,M., Pruss,D. *et al.* (2010) The genome of the domesticated apple (Malus x domestica Borkh.). *Nat. Genet.*, **42**, 833–839.
58. Shulaev,V., Sargent,D.J., Crowhurst,R.N., Mockler,T.C., Folkerts,O., Delcher,A.L., Jaiswal,P., Mockaitis,K., Liston,A., Mane,S.P. *et al.* (2011) The genome of woodland strawberry (Fragaria vesca). *Nat. Genet.*, **43**, 109–116.
59. International Peach Genome Initiative, Verde,I., Abbott,A.G., Scalabrin,S., Jung,S., Shu,S., Marroni,F., Zhebentyayeva,T., Dettori,M.T., Grimwood,J. *et al.* (2013) The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.*, **45**, 487–494.
60. Xu,Q., Chen,L.L., Ruan,X., Chen,D., Zhu,A., Chen,C., Bertrand,D., Jiao,W.B., Hao,B.H., Lyon,M.P. *et al.* (2013) The draft genome of sweet orange (Citrus sinensis). *Nat. Genet.*, **45**, 59–66.
61. Tuskan,G.A., Difazio,S., Jansson,S., Bohlmann,J., Grigoriev,I., Hellsten,U., Putnam,N., Ralph,S., Rombauts,S., Salamov,A. *et al.* (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science*, **313**, 1596–1604.
62. Banks,J.A., Nishiyama,T., Hasebe,M., Bowman,J.L., Gribskov,M., dePamphilis,C., Albert,V.A., Aono,N., Aoyama,T., Ambrose,B.A. *et al.* (2011) The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. *Science*, **332**, 960–963.
63. Tomato Genome Consortium. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635–641.
64. Potato Genome Sequencing Consortium, Xu,X., Pan,S., Cheng,S., Zhang,B., Mu,D., Ni,P., Zhang,G., Yang,S., Li,R. *et al.* (2011) Genome sequence and analysis of the tuber crop potato. *Nature*, **475**, 189–195.
65. Jaillon,O., Aury,J.M., Noel,B., Policriti,A., Clepet,C., Casagrande,A., Choisne,N., Aubourg,S., Vitulo,N., Jubin,C. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.
66. Prochnik,S.E., Umen,J., Nedelcu,A.M., Hallmann,A., Miller,S.M., Nishii,I., Ferris,P., Kuo,A., Mitros,T., Fritz-Laylin,L.K. *et al.* (2010) Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri. *Science*, **329**, 223–226.