Database tool

**GrTEdb: the first web-based database of transposable elements in cotton (Gossypium raimondii)**

Zhenzhen Xu¹†, Jing Liu²†, Wanchao Ni³, Zhen Peng², Yue Guo², Wuwei Ye³, Fang Huang¹, Xianggui Zhang¹, Peng Xu¹, Qi Guo¹, Xinlian Shen¹,⁎ and Jianchang Du²,⁎

¹Key Laboratory of Cotton and Rapeseed (Nanjing), The Institute of Industrial Crops, Jiangsu Academy of Agricultural Sciences, Nanjing 210014, China, ² Provincial Key Laboratory of Agrobiology, The Institute of Biotechnology, Jiangsu Academy of Agricultural Sciences, Nanjing 210014, China and ³State Key Laboratory of Cotton Biology, The Institute of Cotton Research, Chinese Academy of Agricultural Sciences, Anyang 455000, China

⁎Corresponding author: Tel: +86 2584392767; Email: dujianchang@hotmail.com

Correspondence may also be addressed to Xinlian Shen. Tel: +86 2584390291; Email: xlshen68@126.com

†These authors contributed equally to this work.

Citation details: Xu, Z., Liu, J., Ni, W. et al. GrTEdb: the first web-based database of transposable elements in cotton (Gossypium raimondii). Database (2017) Vol. 2017: article ID bax013; doi:10.1093/database/bax013

Received 27 July 2016; Revised 10 January 2017; Accepted 13 January 2017

**Abstract**

Although several diploid and tetraploid Gossypium species genomes have been sequenced, the well annotated web-based transposable elements (TEs) database is lacking. To better understand the roles of TEs in structural, functional and evolutionary dynamics of the cotton genome, a comprehensive, specific, and user-friendly web-based database, *Gossypium raimondii* transposable elements database (GrTEdb), was constructed. A total of 14 332 TEs were structurally annotated and clearly categorized in *G. raimondii* genome, and these elements have been classified into seven distinct superfamilies based on the order of protein-coding domains, structures and/or sequence similarity, including 2929 Copia-like elements, 10 368 Gypsy-like elements, 299 L1, 12 Mutators, 435 PIF-Harbingers, 275 CACTAs and 14 Helitrons. Meanwhile, the web-based sequence browsing, searching, downloading and blast tool were implemented to help users easily and effectively to annotate the TEs or TE fragments in genomic sequences from *G. raimondii* and other closely related *Gossypium* species. GrTEdb provides resources and information related with TEs in *G. raimondii*, and will facilitate gene and genome analyses within or across *Gossypium* species, evaluating the impact of TEs on their host genomes, and investigating the potential interaction between TEs and protein-coding genes in *Gossypium* species.

**Database URL:** http://www.grtedb.org/
Introduction

Transposable elements (TEs) are the most abundant DNA components in most characterized genomes of high eukaryotes (1). Based on their structural features and transposition mechanisms, TEs are generally classified into two classes: retrotransposons and DNA transposons (2). In plants, retrotransposons are further classified into two distinct orders, long terminal repeat (LTR)-retrotransposons (Tyl/Copia and Ty3/Gypsy) and non-LTR retrotransposons (LINE and SINE), whereas DNA transposons are traditionally separated into two main orders, terminal inverted repeat (TIR) (Tc1-Mariner, hAT, Mutator, PIF/Harbinger and CACTA) and Helitron (Helitron) (2, 3). Although TEs are often considered as ‘junk DNA’ due to their continuous reproduction and potential disruption of the regular host genes (4–6), more evidence has unambiguously shown that they play important roles in altering gene structures, regulation of gene expression, affecting genome evolution and creating new genes (7–9). Thus, complete identification and characterization of TEs have become a priority in genome sequencing projects, and this will largely contribute to accurate annotation of protein-coding genes and other genomic components, and play significant roles in investigating potential interaction between TEs and functional genes (10).

Recently, several diploid and tetraploid Gossypium species genomes have been sequenced (11–15), and the availability of their draft genome sequences have provided an unprecedented opportunity for identification, structural and functional characterization and evolutionary analyses of TEs in this economically important crop. Gossypium raimondii (DD; 2n = 6), one of the putative D-genome parent of tetraploid cotton species [such as G. hirsutum (L.) and G. barbadense (L.)] has a smaller genome size (~737.8 Mb) (12). So, we carried out the characterization of almost all families of TEs in G. raimondii genome using comprehensive methods, and constructed the G. raimondii transposable elements database (GrTEdb) in this study. We implemented web-based sequence browsing, searching, downloading and blast tool to help users easily and effectively to annotate the TEs or TE fragments in genomic sequences from G. raimondii and other closely related Gossypium species. Thus, GrTEdb provide the first web-based friendly user interface database of TEs in Gossypium species, and will also facilitate genome evolution analyses within or across Gossypium species, evaluating the impact of TEs on their host genomes, and investigating the potential interaction between TEs and protein-coding genes.

Construction and content of the database

The assembled sequence of the G. raimondii genome was downloaded from http://www.phytozome.com (11).

Table 1. Summary of the identified TEs in G. raimondii

| Class          | Order | Superfamily | Copy numbers |
|----------------|-------|-------------|--------------|
| Retrotransposons | LTR   | Copia       | 2929         |
|                |       | Gypsy       | 10 368       |
| LINE           |       | L1          | 299          |
| DNA transposons | TIR   | Mutator     | 12           |
|                |       | PIF-Harbingers | 435      |
|                |       | CACTA       | 275          |
| Helitron       |       | Helitron    | 14           |
| Total          |       |             | 14 332       |

A combination of structure-based and homology-based approaches was employed to identify different TEs in the G. raimondii genome. LTR-retrotransposons were characterized according to the methods previously described by Ma et al. (2006) (16): first, the LTR-retrotransposons were identified using the LTR_STRUC software; then CROSS_MATCH was used to detect elements missed by the program. The alignments were performed between G. raimondii genome and the flanking LTRs of these LTR-retrotransposons, which generated by the LTR_STRUC. Different perl scripts were written to facilitate the data mining and analyses. Other Non-LTR-retrotransposons and DNA transposons (such as L1, Mutator, PIF-Harbingers, CACTA and Helitron) were detected following the protocol provided by Holligan et al. (2006) (17): the alignment were performed between the conservative sequences of transposase in Arabidopsis thaliana and G. raimondii genomes using tblastn, and the TSD and TIR were detected using some perl scripts. The detailed manual inspection was conducted to confirm each predicted element and to define its structure and boundaries. In addition, TEs were classified into different superfamilies and families as previously described (2, 17). Only elements with clearly defined boundaries and insertion sites were deposited in the GrTEdb database.

Based the above approaches, 14 332 TEs were structurally annotated and clearly categorized in the G. raimondii genome, and these elements are classified into seven distinct superfamilies based on the order of protein-coding domains, structures and/or sequence similarity, including 2929 Copia-like elements, 10 368 Gypsy-like elements, 299 L1, 12 Mutators, 435 PIF-Harbingers, 275 CACTAs and 14 Helitrons (Table 1). Based on the 80-80-80 rule (2). TEs that were assigned as Copia- and Gypsy-like elements superfamilies were then categorized into 199 and 218 distinct families respectively because of their large number in G. raimondii.

User interface

GrTEdb was established to enable users to browse, search, view, analyze and download the TEs data and information.
The GrTEdb database organization is navigated by six sections: Home, Browse, Search and Download, Blast, Links and Contact (Figure 1A).

**Browse**

In the browsing interface, the classification structures of TEs deposited in GrTEdb were showed. Users can download the whole TEs sequences, and can browse any one superfamily of interest by the hyperlinks provided. The detailed information of each superfamily can be retrieved and downloaded by clicking the corresponding entry (Figure 1B).

**Search and download**

In the searching and downloading interface, users can use a keyword to search the GrTEdb (e.g. TE ID, Class, Order, Superfamily and Family) to locate specific TEs quickly. The search results can be viewed and downloaded by clicking the hyperlinks provided on the page (Figure 2).

In the chromosomal region search page, users can retrieve the TEs for any one entire chromosome or in a defined window around either a chromosomal position or a gene model, and the detailed information of each retrieved TEs can be viewed and downloaded by clicking the hyperlinks provided on the page (Figure 3). This function can help users to locate TEs that surround the genes of interests easily, and study the interaction between TEs and their adjacent genes.

**Blast**

We did not intend to integrate tools currently available (except for BLAST) for sequence comparison, editing and/or assembly in our database because of the complex structural variation and distribution patterns of TEs among classes and families (Figure 4). In the BLAST search page, users can handy and quickly compare their sequences with the cotton TEs deposited in GrTEdb.

**Links**

A variety of links to other TEs database were included in our GrTEdb database.

**Contact**

In this section, contact information and links to our labs were provided. Please feel free to contact us if you have any suggestions and problems.

**Discussion**

Because of the structural complexity and the time consuming process, it remains challenging to annotate all TEs in a
sequenced genome. Currently only a few TE databases have been established (10, 18–24). Because these databases can help users easily and quickly annotate their sequences, and they have been widely used (10). However, in these plant TE databases such as P-MITE (a Plant MITE database), the TIGR Plant Repeat Databases, and so on, there is little information about the cotton TEs. In parallel, although there were some reports associated with TEs in Gossypium (11–15, 25, 26), the web-based database of TEs was lacked. Here we have generated a web-based TE database (GrTEdb) using multiple methods, and only TEs with clearly defined boundaries were deposited in the database. More studies have showed that many TEs are structurally incomplete because they have undergone intra- or inter-element unequal recombination or accumulation of small deletions by illegitimate recombination (27, 28). For example, a large number of LTR-RT families with highly degraded protein-coding sequences or without any coding sequences (often defined as non-autonomous elements) have been found in several plants (29–35), and these elements remains challenging to be identified and characterized. Therefore, GrTEdb provides the reference sequences of TEs data for cotton, and users can use these data to identify more complex elements and develop their specific functions.

Recently, G. arboretum (A2) genome, a putative contributor of the A subgenomes cotton species, and the allo-tetraploid upland cotton (AD1 [G. hirsutum (L.)], which accounts for >90% of cultivated cotton worldwide, have been sequenced and assembled (13–15). Because of the

Figure 2. The searching interface of GrTEdb. Users can use a keyword to locate specific TEs quickly in GrTEdb (e.g. TE ID, Class, Order, Superfamily and Family). The search results can be viewed and downloaded by clicking the hyperlinks provided on the page.
close evolutionary relationships of DD, AA and AADD genomes, our GrTEdb database is not only useful for *G. raimondii* study, but also can facilitate structural and evolutionary analysis in AA, DD, AADD and other unfinished *Gossypium* genomes. The web-based interface can also help users at the beginning stage of bioinformatics to easily access and use this database. Further, TEs in our database will help cotton breeders develop markers for mapping agronomically important genes and accelerate breeding process.

**Conclusions**

We have generated a web-based GrTEdb, and it provides researchers with not only resources and information related to different TEs in the cotton genome but also tools
for performing data analysis. Thus, GrTEdb will facilitate cotton genome evolution analyses among AA, DD and AADD genome species, the evaluating impact of TEs on their host genomes, and investigating the potential interaction between TEs and protein-coding genes. In parallel, TEs in our database will facilitate users for marker development for mapping agronomically important genes, and for both intra- and inter-specific comparison of TEs at whole genome levels.

Availability and requirements

All TEs or subsets of TEs can be viewed and downloaded from the website http://www.grtedb.org/, and all data deposited in the database are freely available to all users without any restrictions.

Funding

The Key Scientific and Technological Project of Jiangsu Province (BK20150540); Jiangsu Agricultural Science and Technology Innovation Fund (CX(14)5008); the State Key Laboratory of Cotton Biology Open Fund (CB2016B03); the National Natural Science Foundation of China (NSFC) (31370266, 31471545).

Conflict of interest. None declared.

References

1. Finnegan,D.J. (1985) Transposable elements in eukaryotes. Int. Rev. Cytol., 93, 281–326.
2. Wicker,T., Sabot,F., Hua-Van,A. et al. (2007) A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet., 8, 973–982.
3. Feschotte,C. and Pritham,E.J. (2007) DNA transposons and the evolution of eukaryotic genomes. Annu. Rev. Genet., 41, 331–368.
4. Doolittle,W.F. and Sapienza,C. (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature, 284, 601–603.
5. Ma,J., and Bennetzen,J.L. (2006) Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. Proc. Natl. Acad. Sci. USA, 103, 383–388.
6. Zhang,W., Lee,H.R., Koo,D.H. et al. (2008) Epigenetic modification of centromeric chromatin: hypomethylation of DNA sequences in the CENH3-associated chromatin in Arabidopsis thalana and maize. Plant Cell, 20, 25–34.
7. Bennetzen,J.L. (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. Curr. Opin. Genet. Dev., 15, 621–627.
8. Feschotte,C. (2008) Transposable elements and the evolution of regulatory networks. Nat. Rev. Genet., 9, 397–405.
9. Bucher,E., Reinders,J., and Mirouze,M. (2012) Epigenetic control of transposon transcription and mobility in Arabidopsis. Curr. Opin. Plant Biol., 15, 503–510.
10. Du,J.C., David,G., Tian,Z.X. et al. (2010) SoyTEdb: a comprehensive database of transposable elements in the soybean genome. BMC Genomics, 11, 113.
11. Paterson,A.H., Wendel,J.F., Gundlach,H. et al. (2012) Repeated polyploidization of Gossypium genomes and the evolution of spinable cotton fibres. Nature, 492, 423–427.
12. Wang,K.B., Wang,Z.W., Li,F.G. et al. (2012) The draft genome of a diploid cotton Gossypium raimondii. Nat. Genet., 44, 1098–1103.
13. Li,F.G., Fan,G.Y., Wang,K.B. et al. (2014) Genome sequence of the cultivated cotton *Gossypium arboretum*. Nat. Genet., 46, 567–572.
14. Zhang,T.Z., Hu,Y., Jiang,W.K. et al. (2015) Sequencing of allo-tetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. Nat. Biotechnol., 33, 531–537.
15. Li,F.G., Fan,G.Y., Lu,C.R. et al. (2015) Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. Nat. Biotechnol., 33, 524–530.
16. Ma,J.X., and Jackson,S.A. (2006) Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. *Genome Res.*, 16, 251–259.
17. Holligan,D., Zhang,X., Jiang,N. et al. (2006) The transposable element landscape of the model legume *Lotus japonicus*. Genetics, 174, 2215–2228.
18. Ma,B., Li,T., Xiang,Z.H. et al. (2015) MnTEDb, a collective resource for mulberry transposable elements. *Database*, 2015, 1–10.
19. Chaparro,C., Guyot,R., Zuccolo,A. et al. (2007) RetrOryza: a database of the rice LTR-retrotransposons. *Nucleic Acids Res.*, 35, 66–70.
20. Jurka,J., Kapitonov,V.V., Pavlicek,A. et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, 110, 462–467.
21. 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, 360–363.
22. Murukarthick,J., Sampath,P., Lee,S.C. et al. (2014) BrassicaTED - a public database for utilization of miniature transposable elements in Brassica species. *BMC Res. Notes*, 7, 379.
23. 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.
24. Li,S.F., Zhang,G.J., Zhang,X.J. et al. (2016) DPTEdb, an integrative database of transposable elements in dioecious plants. *Database*, 10.1093/database/baw078.
25. Jennifer,S.H., HyeRan,K., John,D.N. et al. (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Res.*, 16, 1252–1261.
26. Hu,G., Hawkins,J.S., Grover,C.E. et al. (2010) The history and disposition of transposable elements in polyploid *Gossypium*. *Genome*, 53, 599–607.
27. Devos,K.M., Brown,J.K. and Bennetzen,J.L. (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.*, 12, 1075–1079.
28. Ma,J.X., Devos,K.M., and Bennetzen,J.L. (2004) Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res.*, 14, 860–869.
29. Jiang,N., Bao,Z., Temnykh,S. et al. (2002) Dasheng: A recently amplified nonautonomous long terminal repeat element that is a major component of pericentromeric regions in rice. *Genetics*, 161, 1293–1305.
30. Jiang,N., Jordan,I.K. and Wessler,S.R. (2002) Dasheng and RIRE2. A nonautonomous long terminal repeat element and its putative autonomous partner in the rice genome. *Plant Physiol.*, 130, 1697–1705.
31. Kalender,R., Vicent,C.M., Peleg,O. et al. (2004) Large retrotransposon derivatives: Abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics*, 166, 1437–1450.
32. Kejnovsky,E., Kubat,Z., Macas,J. et al. (2006) Retand: a novel family of gypsy-like retrotransposons harboring an amplified tandem repeat. *Mol. Genet. Genomics*, 276, 254–263.
33. Du,J.C., Tian,Z.X., Hans,C.S. et al. (2010) Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multispecific comparison. *Plant J.*, 63, 584–559.
34. Du,J.C., Tian,Z.X., Bowen,N.J. et al. (2010) Bifurcation and enhancement of autonomous-nonautonomous retrotransposon partnership through LTR swapping in soybean. *Plant Cell*, 22, 48–61.
35. Zhao,M.X., and Ma,J.X. (2013) Co-evolution of plant LTR-retrotransposons and their host genomes. *Protein Cell*, 4, 493–501.