Journal of Cotton Research

COMMENT

Open Access

Towards complete deconstruction of cotton transcriptional landscape

LI Fuguang

Abstract

Recently, Wang et al. systematically explored the transcription landscape in diploid cotton *Gossypium arboreum*. In the study, they integrated four high-throughput sequencing techniques, including Pacbio sequencing, strand-specific RNA sequencing (ssRNA-seq), Cap analysis gene expression sequencing (CAGE-seq), and PolyA sequencing (PolyA-seq) to profile the RNA transcriptome of *G. arboreum*. They developed a pipeline, IGIA to construct accurate gene structure annotation based on the updated genome of *G. arboreum* and the multi-strategic RNA-seq data. Their study revealed some intriguing phenomena and potential novel mechanisms in the regulation of RNA transcription in plants, and also provided valuable resources for further functional genomic research in cotton.

Keywords: Cotton, CAGE-seq, PolyA-seq, Pacbio-seq, Transcriptome, *Gossypium arboreum*, IGIA, MagenDB

Main text

Cotton is an important natural fiber crop, and also a kind of model plant for studying cell differentiation, elongation, and cell wall development. From 2012 to now, four different cotton genomes including allotetraploid (AADD) *G. hirsutum* (Li et al. 2015; Zhang et al. 2015; Hu et al. 2019; Wang et al., 2019c; Yang et al. 2019) and *G. barbadense* (Yuan et al. 2016; Liu et al. 2015; Hu et al. 2019; Wang et al. 2019) genomes, and two ancestors diploid *G. raimondii* (DD) (Wang et al. 2012; Paterson et al. 2012) and *G. arboretum* (AA) (Li et al. 2014; Du et al. 2018), were sequenced and updated. However, all of the genomes still lack comprehensive and high-resolution gene annotation maps, which would undoubtedly hinder deeper study on functional genome in cotton.

CGAP and IGIA pipeline

In order to eliminate this limitation, Zhu's lab started a cotton genome annotation project (CGAP) which aims to decode the transcriptome of four cotton genomes. A recent study in *Nature Communications* (Wang et al. 2019b) reported their first achievement in genome annotations for Asian cotton (*G. aboreum*), which was a cultivar once widely cultivated in the Asian region. *G. aboreum* is one of the ancestral genome donor species of tetraploid cultivated cotton, *G. hirsutum* and *G. barbadense*. The haploid genome of Asian cotton consists of about 1.7G nt and contains 40 960 protein-coding genes (Du et al. 2018). Wang et al. applied four complementary high-throughput technologies including Pacbio iso-seq for full-length transcript sequence information, strand-specific ssRNA-seq for high-depth quantitative expression and splicing information, and CAGE-seq and PolyA-seq for accurate information on gene transcription initiation sites (TSSs) and termination sites (TESs) (Wang et al. 2019b). They developed IGIA software to integrate high-depth sequencing data: Pacbio-seq 92.8Gb, ssRNA-seq 455.1Gb, CAGE-seq 158.2Gb, and PolyA-seq 339.2Gb.

RNA splicing, microexon and AS hotspot

The study found that 17 101 genes (43.3% of the total expressed genes in 16 tissues) had more than two transcriptional isomers caused by alternative splicing. By integrating the results of NGS (next generation sequencing) and PacBio sequencing in this study, it was found that the splicing junctions predicted only based on the PacBio sequencing results contain a large number of bubble errors and boundary errors. This indicated that the number of

Correspondence: lifuguang@caas.cn
Institute of Cotton Research of Chinese Academy of Agricultural Sciences, Anyang 45500, China

gene isoforms predicted based on the PacBio full-length transcript sequencing in the previous studies would be significantly overestimated.

Microexon, as short as 3 nt in length, was first reported in animals. Wang et al.'s study was a pilot report to identify microexons in plants. In a comparison of multiple plant species, a 45 nt conserved microexon with a potentially important role was found. It is located in the middle region of the transcription factor *AP2* gene. Surprisingly, the length of this microexon remains unusual stability from gymnosperms to flowering plants, implying that the microexon might exert conserved function during evolution. Wang et al. verified that the switch of the microexon could adjust binding affinity to its DNA targets by using EMSA assay.

Interestingly, they found that some local regions of cotton genes exhibit high frequency of alternative splicing (AS), which was named as AS hotspot. By comparing the transcriptome annotation data of multiple plants and animals, they found that the AS hotspot may be a common AS phenomenon in plants, but not in animals. This special phenomenon deserves further study.

TSS/TES switch and their functions

With high-accuracy and high-throughput RNA 5′ and 3′ ends identification technologies: CAGE-seq and PolyA-seq, Wang et al. obtained the TSS and TES information of the Asian cotton genome. They used the annotation information in TAIR database (v11) to analyze the enrichment degree of feature motifs around the TSS and TES. The results showed that their annotation quality was significantly better than *Arabidopsis* in terms of the accuracy of TSS and TES annotation.

By analyzing 44 728 TSS clusters from 22 863 identified genes in Asian cotton, they found that 38.4% had two or more TSSs. Among the TSSs for a multiple-TSS gene, distal TSS is always used more frequently than the proximal TSS. They also revealed that alternative promoter usage could alter the length of 5′UTR or N-terminus of protein for 5 888 and 2 800 genes, respectively. Their study also found that alternative promoter usage of multiple-TSS genes, such as *NRT1.2*, *REC1*, *LRR*, and *PP2C*, is specific to the ovule developmental stage. During the ovule development, the TSS switch for a nitrate transporter, *NRT1.2* would cause the loss of the four transmembrane domains of the NRT1.2 protein, resulting in producing two protein isomers, NRT-L and NRT-S. Based on 3D homology modeling, they speculated that the 3D structure of NRT-S was significantly looser in transmembrane region than that of NRT-L. Moreover, they also used the transporter assay in HEK-293 cells to compare their transport capacity of nitrate and found that the transporting ability of NRT-S decreased significantly.

In addition, they also analyzed the TESs based on the PolyA-seq. The pervasive alternative polyadenylation events were also identified during the tissue development or differentiation process in cotton. In particularly, they observed a gradual shortening of 3′ UTR for all expressed genes during ovule development (0 to 20 DPA) and a sudden reversal to lengthen the 3′ UTR at mature seed, the finally developmental stage of ovule. However, what the function and mechanism of this lengthening transition need to be further clarified in the future.

First discovery of polycistron in cotton

Finally, Wang et al. analyzed the full-length transcripts of PacBio sequencing, and for the first time, they discovered that approximately 5% of genes in Asian cotton took place transcription read-through, producing polycistron transcripts similar to those of prokaryotes. These genes in polycistrons are close to each other, and their average distance is significantly shorter than other adjacent independent genes. In addition, the functional analysis of the gene pairs in the polycistrons showed that they tend to either have the same function or be located in the same molecular network/pathway.

Generally, the genes in eukaryotes are thought to transcribe to transcript with a main open reading frame (monocistron). For the first time, Wang et al.'s study found that polycistronic transcription is widespread in eukaryotic genomes. This will be a new discovery with the potential to rewrite textbooks.

Prospective and MagenDB database

Wang et al.'s study used a systematic and comprehensive transcription analysis to explain the panorama of Asian cotton genome transcription. Their study revealed some novel transcription phenomena and mechanisms in the regulation of RNA transcription in plants, and opened new research horizons for researchers in the field of RNA transcription regulation. Meanwhile, it provided valuable resources for cotton community. To facilitate user to query, browse, and compare these data, they integrate them into the newly developed database MagenDB (http://magen.whu.edu.cn/magendb/)(Wang et al. 2019a).

Authors' contributions
Li FG prepared and wrote the manuscript. The author read and approved the final manuscript.

Funding
No Funding.

Availability of data and materials
Not applicable.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 13 February 2020 Accepted: 12 March 2020
Published online: 05 April 2020

References

Du X, Huang G, He S, et al. Resequencing of 243 diploid cotton accessions based on an updated a genome identifies the genetic basis of key agronomic traits. Nat Genet. 2018;50:796–802. https://doi.org/10.1038/s41588-018-0116-x.

Hu Y, Chen J, Fang L, et al. *Gossypium barbadense* and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton. Nat Genet. 2019;51:739–48. https://doi.org/10.1038/s41588-019-0371-5.

Li F, Fan G, Lu C, et al. Genome sequence of cultivated upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. Nat Biotechnol. 2015;33:524–30. https://doi.org/10.1038/nbt.3208.

Li F, Fan G, Wang K, et al. Genome sequence of the cultivated cotton *Gossypium arboreum*. Nat Genet. 2014;46:567–72. https://doi.org/10.1038/ng.2987.

Liu X, Zhao B, Zheng HJ, et al. *Gossypium barbadense* genome sequence provides insight into the evolution of extra-long staple fiber and specialized metabolites. Sci Rep. 2015;5:14139. https://doi.org/10.1038/srep14139 .

Paterson AH, Wendel JF, Gundlach H, et al. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature. 2012;492:423–7. https://doi.org/10.1038/nature11798 .

Wang D, Fan W, Guo X, et al. MaGenDB: a functional genomics hub for Malvaceae plants. Nucleic Acids Res. 2019a;48(D1):D1076–84. https://doi.org/10.1093/nar/gkz953 .

Wang K, Wang D, Zheng X, et al. Multi-strategic RNA-seq analysis reveals a high-resolution transcriptional landscape in cotton. Nat Commun. 2019b;10:4714. https://doi.org/10.1038/s41467-019-12575-x.

Wang K, Wang Z, Li F, et al. The draft genome of a diploid cotton *Gossypium raimondii*. Nat Genet. 2012;44:1098–103. https://doi.org/10.1038/ng.2371.

Wang M, Tu L, Yuan D, et al. Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. Nat Genet. 2019c;51:224–9. https://doi.org/10.1038/s41588-018-0282-x .

Yang Z, Ge X, Yang Z, et al. Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. Nat Commun. 2019;10. https://doi.org/10.1038/s41467-019-10820-x.

Yuan D, Tang Z, Wang M, et al. The genome sequence of Sea-island cotton (*Gossypium barbadense*) provides insights into the allopolyploidization and development of superior spinnable fibres. Sci Rep. 2016;5:17662. https://doi.org/10.1038/srep17662 .

Zhang T, Hu Y, Jiang W, et al. Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. Nat Biotechnol. 2015;33:531–7. https://doi.org/10.1038/nbt.3207.