

COMMENT

Open Access

# Unraveling the puzzle of the origin and evolution of cotton A-genome



MA Zhiying

## Abstract

*Gossypium hirsutum*, the most widely planted cotton species, its evolution has long been an unsolved puzzle because of its hybrid origin from D-genome and A-genome species. To better understand the genetic component of cotton, Huang et al. recently sequenced and assembled the first  $A_1$ -genome *G. herbaceum*, and updated the  $A_2$ -genome *G. arboreum* and  $(AD)_1$ -genome *G. hirsutum*. On the basis of the three reference genomes, they resolved existing controversial concepts and provided novel evolutionary insights surrounding the A-genome.

**Keywords:** Cotton genomes, *Gossypium herbaceum*, Origins, Evolution, Structural variations

## Main text

Cotton is one of the most important economic crops in the world, and more than 90% of the cotton fiber is derived from the  $(AD)_1$ -genome Upland cotton (*G. hirsutum*) (Ma et al. 2018). The  $(AD)_1$ -genome is formed by the natural hybridization between D- and A-genome (Li et al. 2015). *G. herbaceum* ( $A_1$ -genome) and *G. arboreum* ( $A_2$ -genome) are the only two extant diploid A genomes. Abundant studies supported that the  $D_5$ -genome *G. raimondii* could be regarded as the D-subgenome donor of  $(AD)_1$ -genome (Wendel 1989). However, as commented in *Science* (Zahn 2012), the A-subgenome donor in tetraploid genome has been an unsolved mystery and attracts a great research interest. Previously, either  $A_1$ - or  $A_2$ -genome has been reported to be the actual A-genome donor of tetraploid cotton.

## First $A_1$ genome sequence and high-quality genome updates of $A_2$ and $(AD)_1$

Since the release of the first  $D_5$ -genome in 2012 (Wang et al. 2012), the genome of the second  $D_5$  (Paterson et al. 2012; Udall et al. 2019), and that of  $A_2$  (Du et al. 2018; Li et al. 2014),  $(AD)_1$  (Hu et al. 2019; Li et al. 2015; Wang et al. 2019; Yang et al. 2019; Zhang et al.

2015) as well as  $(AD)_2$  (a much less cultivated *G. barbadense*) (Hu et al. 2019; Liu et al. 2015; Wang et al. 2019; Yuan et al. 2015), were sequenced and assembled with great updating efforts resulted in much improvements for these genomes. Recently, Huang et al. sequenced and assembled  $A_1$  variety *africanum* for the first time and re-assembled high-quality  $A_2$  cultivar Shixiya1 and  $(AD)_1$  genetic standard Texas Marker-1 (TM-1) genomes on the basis of PacBio long reads, paired-end sequencing and high-throughput chromosome conformation capture (Hi-C) technologies. The resulting  $A_1$  assembly captured 1 556 megabases (Mb) of genome sequences and 95.69% of total genome sequences spanning 1 489 Mb were categorized and ordered into 13 pseudo-chromosomes. They substantially improved the existing  $A_2$  and  $(AD)_1$  genome assemblies, mainly from aspects of genome completeness and accuracy (Huang et al. 2020). These updated  $A_2$ - and  $(AD)_1$ -genomes will certainly replenish earlier assemblies as chromosome-scale references. With high-quality assembly of the three cotton species, a more complete landscape of genome architecture, gene annotations and transposable element (TE) insertions were provided, which is critical to evolutionary and comparative genomics as well as genetic variation analysis.

Correspondence: [mzhy@hebau.edu.cn](mailto:mzhy@hebau.edu.cn)

State Key Laboratory of North China Crop Improvement and Regulation, Hebei Agricultural University, Baoding 071001, China



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## Origin of allotetraploid cotton and two diploid A-genomes

On the basis of multiple evidences, such as molecular tree, whole-genome phylogenetic relationships and population analysis, the authors propose that all existing A-genomes may have originated from a common ancestor, referred to here as  $A_0$  (Huang et al. 2020). The ancient and extinct  $A_0$  lineage is likely the ancestral form of one of the original species parents of  $(AD)_1$  rather than the extant  $A_1$  and  $A_2$  lineages. Relatively,  $A_0$  was more phylogenetically related to the present  $A_1$  than  $A_2$ . Hybridization of this  $A_0$  genome with a  $D_5$  species eventually produced the current allotetraploid cotton at ~1.6 million years ago (MYA), which preceded the speciation of current  $A_1$  and  $A_2$  at ~0.7 MYA. This genome-based analysis will likely close down the  $A_1$  versus  $A_2$  argument, especially if archaeological data is dugged up someday to pinpoint out the existence of the currently-extinct  $A_0$  genome. Since the publication of these new findings, our knowledge of cotton genome evolution will have to be revised and updated.

## Developing novel methods to detect genome expansion and evolution

The authors suggest that cotton genome evolution is characterized by bursts of transposable element (TE) activity. The two A-genomes and the A-subgenome of  $(AD)_1$  experienced expansion in genome size that was highly correlated with TE bursts. As much as 72.57% of the  $A_1$ -genome and 73.62% of the  $A_2$ -genome were composed of LTR-type TEs. By using fragmented coding sequences of LTR-type TEs, they developed a novel method, named Gaussian probability density function (GPDF) analysis, to overcome a major pitfall related to traditional method that relied on the presence of both ends of full-length LTRs, such that more recently inserted LTRs are likely over-represented. They suggested that several long-terminal-repeat bursts which occurred from 5.7 MYA to less than 0.61 MYA contributed compellingly to A-genome size expansion, speciation and evolution.

## Characterization of structural variations related to cotton fiber development

The authors characterized structural variations (SVs) in cotton genomes using comparative genomic analysis, which provided dozens of putative candidate genes to investigate phenotypic differences among the three cotton species. Abundant species-specific SVs in genic regions changed the expression of many important genes, some of which were found to alter expression intensities of genes involved in fatty acid biosynthesis, leading probably to spinnable fiber improvement in  $(AD)_1$  compared with  $A_1$  or  $A_2$ , as confirmed by several transgenic cotton lines that overexpressed *KCS6* gene.

In conclusion, Huang and colleagues assembled three high-quality cotton reference genomes, and performed systematic and comprehensive genome analysis, especially for cotton A-genomes, which represents a major step toward understanding the evolution of the cotton A genome. They not only have provided the scientific community with valuable genomic and genetic resources to facilitate genetic evolution, comparative genomic analysis, but also will accelerate the process of cotton genetic improvement with advanced methodology and new varieties.

### Acknowledgements

We thank editorial board for inviting us to write comments on this interesting work and thank them for their suggestions in the manuscript revision.

### Author's contributions

Ma ZY wrote the manuscript. The author(s) read and approved the final manuscript.

### Funding

No Funding.

### Availability of data and materials

No other data related to this study is available at this time.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The author declares no competing interests.

Received: 28 April 2020 Accepted: 11 May 2020

Published online: 02 June 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>.
- Huang G, Wu Z, Percy RG, et al. Genome sequence of *Gossypium herbaceum*, and genome update of *G. arboreum* and *G. hirsutum* provide insights into cotton A-genome evolution. *Nat Genet.* 2020;52:516–24. <https://doi.org/10.1038/s41588-020-0607-4>.
- 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 H, 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>.
- Ma Z, He S, Wang X, et al. Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. *Nat Genet.* 2018;50:803–13. <https://doi.org/10.1038/s41588-018-0119-7>.
- 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>.
- Udall JA, Long E, Hanson C, et al. De Novo genome sequence assemblies of *Gossypium raimondii* and *Gossypium turneri*. *G3.* 2019;9:3079–85. <https://doi.org/10.1534/g3.119.400392>.

- 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. 2019;51:224–9. <https://doi.org/10.1038/s41588-018-0282-x>.
- Wendel JF. New World tetraploid cottons contain Old World cytoplasm. Proc Natl Acad Sci U S A. 1989;86:4132–6. <https://doi.org/10.1073/pnas.86.11.4132>.
- Yang Z, Ge X, Yang Z, et al. Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. Nat Commun. 2019;10:2989. <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>.
- Zahn LM. Unraveling the origin of cotton. Science. 2012;335:1148. <https://doi.org/10.1126/science.335.6073.1148-a>.
- 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>.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

