Improved *Gossypium raimondii* Genome Using a Hi-C-based Proximity-Guided Assembly

Guoli song (✉ sglzms@163.com)

Chinese Academy of Agricultural Sciences Cotton Research Institute <https://orcid.org/0000-0003-3236-9286>

QiuHong Yang

Chinese Academy of Agricultural Sciences Cotton Research Institute

Dongyun Zuo

Chinese Academy of Agricultural Sciences Cotton Research Institute

Hailiang Cheng

Chinese Academy of Agricultural Sciences Cotton Research Institute

Youping Zhang

Chinese Academy of Agricultural Sciences Cotton Research Institute

Qiaolian Wang

Chinese Academy of Agricultural Sciences Cotton Research Institute

Javaria Ashraf

Chinese Academy of Agricultural Sciences Cotton Research Institute

Xiaoxu Feng

Chinese Academy of Agricultural Sciences Cotton Research Institute

Simin Li

Chinese Academy of Agricultural Sciences Cotton Research Institute

Xiaoqin Chen

Chinese Academy of Agricultural Sciences Cotton Research Institute

Shang Liu

Chinese Academy of Agricultural Science Cotton Research Institute

Research

Keywords: *Gossypium raimondii*, Hi-C, genome assembly, heatmap and collinearity

Posted Date: January 22nd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-63005/v3>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Journal of Cotton Research on August 5th, 2021. See the published version at <https://doi.org/10.1186/s42397-021-00096-2>.

Abstract

Introduction: Genome sequence plays an important role in both basic and applied studies. *Gossypium raimondii*, the putative contributor of the D subgenome of Upland cotton (*G. hirsutum*), highlights the need to improve the genome quality rapidly and efficiently.

Methods: We performed Hi-C sequencing of *G. raimondii* and reassembled its genome based on a set of new Hi-C data and previously published scaffolds. We also compared the reassembled genome sequence with the previous published *G. raimondii* genomes for gene and genome sequence collinearity.

Result: A total of 98.42% of scaffold sequence was clustered successfully, among which 99.72% of the clustered sequence was ordered and 99.92% of the ordered sequence was oriented with high-quality. Further evaluation of results by heat-map and collinearity analysis revealed that the current reassembled genome is significantly improved than the previous one (Wang et al. 2012).

Conclusion: This improvement in *G. raimondii* genome not only provides a better reference genome to increase study efficiency but also offers a new way to assemble cotton genomes. Furthermore, Hi-C data of *G. raimondii* may be used for 3D structure research or regulating analysis.

Introduction

Over the last decade, next-generation sequencing (NGS) technologies have brought immense improvements in plant genome sequencing throughput and reduced the cost, and many plant genomes have been sequenced using this technology such as *F. vesca* (Shulaev et al. 2011), *C. cajan* (Varshney et al. 2012), *G. raimondii* (Wang et al. 2012; Paterson et al. 2012; Udall et al. 2019), *G. arboreum* (Li et al. 2014) and *G. hirsutum* (Li et al. 2015; Zhang et al. 2015). However, *de novo* assembly of large eukaryotic genomes has remained a great challenge with the NGS platform due to a significant amount of repeat contents. As a result, *de novo* assembly of chromosome-scale scaffolds has become a major constraint to the completion of a high-quality genome sequence. Compared with the traditional method, like genetic map and physical map, high-throughput chromosome conformation capture (Hi-C) technology has assisted in the assembly of long scaffolds to produce chromosome-scale genome assemblies (Lightfoot et al. 2017). The Hi-C technique is based on existing scaffolds and restriction enzyme cutting sites, which are more evenly distributed and have much higher density.

The Hi-C-based proximity-guided assembly was initially developed to study the three-dimensional (3D) conformation of chromosomes of yeast gene expression (van Berkum et al. 2010). The working principle based on two regular models with Hi-C data: the first one is that the rate of Hi-C interaction is inversely proportional to the genomic distance between the pairs of loci, the second one is that the rate of Hi-C interaction of pairs of loci within a chromosome is significantly higher than that in different chromosomes (Xie et al. 2015). Based on these two models, Hi-C-based proximity-guided assembly was applied for *de novo* assembling of human, and subsequently for the assembling of *mouse* and *Drosophila* genomes which reported good results or improvement (Burton et al. 2013). With the success

of testing and verifying this method in *Arabidopsis thaliana* (Xie et al. 2015), Hi-C based proximity-guide assembly has been reported as an effective and efficient method which subsequently has been used in many other plants.

Besides its commercial value, cotton also serves as a perfect model system for studying cell wall biosynthesis (Zhang et al. 2018), cell elongation (Guo et al. 2017), and polyploidization (Yuan et al. 2015). The *Gossypium* genus comprises more than 50 species including at least 5 tetraploid species and 45 diploid species. Diploid cottons are divided into 8 sub-genomes, denoted A-G and K based on chromosome pairing relationships (Wendel et al. 1992). Tetraploid cottons, such as cultivated *G. hirsutum* (AD₁) and *G. barbadense* (AD₂), had formed by an allopolyploidy event about 1-2 million years ago (Paterson et al. 2012). These tetraploid cotton species share common ancestors with the modern New World species *G. raimondii* (D₅) and the Old World A-genome species *G. herbaceum* (A₁) or *G. arboreum* (A₂). Previously, genomes of different cotton species sequenced and assembled including *G. raimondii* (Wang et al. 2012), *G. arboreum* (Li et al. 2014), and *G. hirsutum* (Li et al. 2015).

Among these cotton species, the genome of *G. raimondii* has the lowest complexity which has been sequenced and assembled using the next-generation Illumina paired-end sequencing strategy (Wang et al. 2012). Approximately 73% (281 scaffolds) of the assembled sequences were anchored to 13 chromosomes, covering 88.1% of the genome, while only 52.4% (228 scaffolds) of the total sequence was both ordered and oriented. The completeness and accuracy of the previous sequenced and assembled genome of *G. raimondii* (Wang et al. 2012) were relatively low due to higher numbers of repeat elements and low numbers of genetic markers. In the present study, we conducted a *de novo* Hi-C sequencing of *G. raimondii*, and incorporated the new Hi-C data with the existing *G. raimondii* scaffolds (Wang et al. 2012) to improve the quality of the D-genome.

Methods

1. Tissue collection and Hi-C sequencing

1.1 Plant materials

The seeds of *G. raimondii* D₅₋₁ were planted in an incubator at constant environmental conditions having 27°C temperature, 60% relative humidity, 16/8-h light/dark photoperiod, and 100% fluorescent light. When the sixth euphylla came out, these seedlings were transplanted into big pots. Approximately 3-gram young leaves from *G. raimondii* plants were collected and immediately treated with formaldehyde.

1.2 Hi-C pipeline

During this study, we have used the same Hi-C pipeline as in *Arabidopsis thaliana* (Xie, Zheng et al. 2015). Before starting this experiment, we have tested the integrity of DNA from the formaldehyde-treated tissue, and then the DNA was isolated and digested by *Mbol* instead of *HindIII* because of the shorter recognition site (only four bases of *Mbol*). The resulting sticky ends were filled with nucleotides in which cytosine is

biotinylated, and ligated the adjacent blunt ends to a chimeric circle under extremely dilute conditions. Subsequently, DNA was purified and broken into 300-500 base pairs using ultrasonic, pull-down the biotin-labeled DNA and performed the PCR reaction (10 cycles). After DNA purification, the finished Hi-C library was sequenced with an Illumina Hiseq (PE150). A total of 570,412,361 read-pairs were obtained.

2. Genome assembly based on Hi-C data

Assembling of *G. raimondii* genome involved three steps. First, valid Hi-C paired-end reads and contact matrix with a resolution of 100 kb were generated by HiC-Pro (Servant, Varoquaux et al. 2015). The raw sequence data with low quality, unmapped and invalid mapped paired reads were filtered out by HiC-Pro and contact-matrix based on interaction frequency was created. HiC-Pro results show that 95.6% of sequences is clean Q30 bases, showing a good quality of sequence data. After the Hi-C data were filtered out, 81.95% of uniquely mapped sequences are valid paired-end reads. Thus, the valid paired-end reads (223,304,666) were used for further genome assembly. At the second step, Errors in scaffolds of the initial draft assembly were identified and corrected following the *Aedes aegypti*'s *de novo* assembly procedure (Dudchenko, Batra et al. 2017). Briefly, the errors were corrected by identifying the bins where a scaffold's long-range contact pattern changes abruptly, which is unlikely for a correct scaffold. We cut out the error bins as a new scaffold. There are 259 errors within scaffolds. At the third step, the *G. raimondii* genome was assembled with the Hi-C data by Lachesis (Burton et al. 2013), which contained clustering, ordering, and orienting. Lastly, the assembled *G. raimondii* genome was assessed by heat-map and collinear analysis.

3 Repeats and gene annotation

Repeat sequences of *G. raimondii* genome were masked by RepeatMasker (v4.0.8; Smit et al. 2019) with a custom library generated from RepeatModeler(v2.0.1; Flynn et al. 2020). Repeat-masked sequences were obtained as a draft genome in gene prediction. *de novo* prediction, homology-based prediction and transcriptome-based prediction were combined to annotate the draft genome of *G. raimondii*. glimmerHMM (v3.0.4; Chrysanthou et al. 2011) and AUGUSTUS(v3.3.2; Stanke et al. 2006) were run for *de novo* prediction, We used 9 transcriptomes (SRR389181, SRR203240, SRR203250, SRR8878792, SRR8878720, SRR8878562, SRR8878553, SRR8878556, SRR8878559) of *G. raimondii* from different organs and growth stage to predict genes; Homologous proteins from arabidopsis, maize, and rice were input into GenomeThreader (v1.6.1; Gremme. 2015) to train models for homology-based prediction. Sequences of transcriptome samples were aligned to draft genome by HISAT2 (v2.1.0; Kim et al. 2019) and transcripts were assembled by StringTie (v1.3.6; Pertea et al. 2015). All files in general feature format were integrated into a final genome annotation file by EvidenceModeler (v1.1.1; Haas et al. 2008).

The reassembled genome was evaluated in three aspects, completeness, consistence, and continuity. We evaluated continuity according to the length of scaffoldN50 calculated by our python script. Completeness was checked by BUSCO(v4.1.4; Simão et al. 2015) and consistency was evaluated by comparative analysis between our reassembled genome and other published genomes of *G. Raimondii*. Genome aligned were performed by minimap2(v2.1; Heng. 2018) and dot plots were generated by

dotPlotly (<https://github.com/tpoorten/dotPlotly/>). Orthologous genes of *G. Raimondii* genomes were detected by Orthofinder(v2.5.1; Emms *et al.* 2015), visualization of orthologous gene detection were implemented by jcvii (<https://github.com/tanghaibao/jcvii>). All parameters of software in this article were set as default.

Results

1 Assembling results

The genome of *G. raimondii* was reassembled using Hi-C data. About 98.42% of total sequence length was clustered successfully, among which 99.72% and 97.37% of the clustered sequence were ordered and high-quality ordered, respectively. And approximately 99.92% of the ordered sequence was oriented with high-quality. The statistics of pseudo-chromosome length are shown in Table 1, while the indicator statistics of the initial draft genome and reassembly results are shown in Tables 2 and 3, respectively. From the parameters like scaffolds number, N50 and N60 of previously draft genome and reassemble genome, we found that the *G. raimondii* genome using a Hi-C-based proximity-guided assembly is much better than the reported draft genome (Wang et al. 2012).

2 genome assessment results

The quality of the reassembled genome could be verified in four aspects. First, BUSCO analysis reported 97.2% (>90%) completeness scores for the reassembled genome, indicating it is complete. Compared to ~32000 genes in the old draft genome (besides scaffolds), 3,8975 genes have been found in the annotation of the reassembled genome. Second, The heat-map directly proves the validity of the processing methods. Based on the two regular models as we mentioned above, the heat-map of the reassembled results revealed that the diagonal interaction is much higher. The boundaries of each pseudo-chromosome are relatively clear and it is under low background noise that shows good reassembling results (Fig. 1). Third, we compared the draft genome and reassembled genome with two previous genome assemblies of *G. raimondii* (Paterson et al. 2012, Udall et al. 2019) (Fig. 2). The reassemble genome had a general agreement in their alignments with two previous genome assemblies than the draft genome. Fourth, we compared homologous genes in the draft genome and reassembled genome with which in two previous genome assemblies of *G. raimondii* (Paterson et al. 2012, Udall et al. 2019) (Fig. 3), it showed a good agreement, and the size of chromosome in the reassembled genome is more believable than the draft genome.

Discussion

In the present study, we applied *de novo* Hi-C sequencing of the *G. raimondii* genome to improve the quality and accuracy of its previously reported draft genome from Wang et al. (2012). Results from the comparative analysis of different parameters between the draft genome (Wang et al. 2012) and the

current reassembled genome showed a significantly improved quality as compared to the previous one. And Hi-C data can be used in a future study as basic data, such as 3D model genome.

Abbreviations

NGS: next generation sequencing

Hi-C: high-throughput chromosome conformation capture

CAAS: Chinese Academy of Agricultural Science

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

Not applicable.

Funding

Not applicable.

Authors' contributions

Yang QH and Song GL conceived and designed the experiments; all authors performed data analysis and interpretation; Yang QH, Javaria A, Liu S and Song GL wrote the manuscript.

Acknowledgements

This work was supported by CAAS.

Data availability:

The reassembled genome sequences and annotation files will be upload after received.

References

Stanke, M. , Keller, O. , Gunduz, I. , Hayes, A. , Waack, S. , & Morgenstern, B. . (2006). B: AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res 2006, 34(Web Server issue):W435-9. Page 10

of 11 (page number not for citation purposes. Bioinformatics.

Belton, J. M., R. P. McCord, J. H. Gibcus, N. Naumova, Y. Zhan and J. Dekker (2012). "Hi-C: a comprehensive technique to capture the conformation of genomes." Methods **58**(3): 268-276.

Bickhart, D. M., B. D. Rosen, S. Koren, B. L. Sayre, A. R. Hastie, S. Chan, J. Lee, E. T. Lam, I. Liachko, S. T. Sullivan, J. N. Burton, H. J. Huson, J. C. Nystrom, C. M. Kelley, J. L. Hutchison, Y. Zhou, J. Sun, A. Crisa, F. A. Ponce de Leon, J. C. Schwartz, J. A. Hammond, G. C. Waldbieser, S. G. Schroeder, G. E. Liu, M. J. Dunham, J. Shendure, T. S. Sonstegard, A. M. Phillippy, C. P. Van Tassell and T. P. Smith (2017). "Single-molecule sequencing and chromatin conformation capture enable *de novo* reference assembly of the domestic goat genome." Nat Genet **49**(4): 643-650.

Burton, J. N., A. Adey, R. P. Patwardhan, R. Qiu, J. O. Kitzman and J. Shendure (2013). "Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions." Nat Biotechnol **31**(12): 1119-1125.

Chrysanthou, N. , Chrysos, G. , Sotiriades, E. , & Papaefstathiou, I. . (2011). Parallel accelerators for GlimmerHMM bioinformatics algorithm. Design, Automation & Test in Europe Conference & Exhibition. IEEE.

Dong, Y., M. Xie, Y. Jiang, N. Xiao, X. Du, W. Zhang, G. Tosser-Klopp, J. Wang, S. Yang, J. Liang, W. Chen, J. Chen, P. Zeng, Y. Hou, C. Bian, S. Pan, Y. Li, X. Liu, W. Wang, B. Servin, B. Sayre, B. Zhu, D. Sweeney, R. Moore, W. Nie, Y. Shen, R. Zhao, G. Zhang, J. Li, T. Faraut, J. Womack, Y. Zhang, J. Kijas, N. Cockett, X. Xu, S. Zhao, J. Wang and W. Wang (2013). "Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). " Nat Biotechnol **31**(2): 135-141.

Dudchenko, O., S. S. Batra, A. D. Omer, S. K. Nyquist, M. Hoeger, N. C. Durand, M. S. Shamim, I. Machol, E. S. Lander, A. P. Aiden and E. L. Aiden (2017). "*De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds." Science **356**(6333): 92-95.

Ekblom, R. and J. B. Wolf (2014). "A field guide to whole-genome sequencing, assembly and annotation." Evol Appl **7**(9): 1026-1042.

Emms, D. M. , & Kelly, S. . (2015). Orthofinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome biology*, 16(157), 157.

Fierst, J. L. (2015). "Using linkage maps to correct and scaffold *de novo* genome assemblies: methods, challenges, and computational tools." Front Genet **6**: 220.

Flynn, J. M. , Hubley, R. , Rosen, J. , Clark, A. G. , & Smit, A. F. . (2020). Repeatmodeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, 117(17), 201921046.

- Guo, K., L. Tu, Y. He, J. Deng, M. Wang, H. Huang, Z. Li and X. Zhang (2017). "Interaction between calcium and potassium modulates elongation rate in cotton fiber cells." J Exp Bot **68**(18): 5161-5175.
- Gremme, G. . (2013). Computational gene structure prediction.
- Haas, B. J. , Salzberg, S. L. , Zhu, W. , & Mihaela Pertea.... (2008). Automated eukaryotic gene structure annotation using evidencemodeler and the program to assemble spliced alignments. *Genome Biology*, 9(1), R7.
- Heng, L.(2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*.
- Jibran, R., H. Dzierzon, N. Bassil, J. M. Bushakra, P. P. Edger, S. Sullivan, C. E. Finn, M. Dossett, K. J. Vining, R. VanBuren, T. C. Mockler, I. Liachko, K. M. Davies, T. M. Foster and D. Chagné (2018). "Chromosome-scale scaffolding of the *black raspberry* (*Rubus occidentalis* L.) genome based on chromatin interaction data." Horticulture Research **5**(1).
- Jonathan F. Wendel, V. A. A. (1992). "Phylogenetics of the cotton genus(*Gossypium* L.): character-state weighted parsimony analysis of chloroplast DNA restriction sites data and its systematic and biogeographic implications " JSTOR.
- Kim, D. , Paggi, J. M. , Park, C. , Bennett, C. , & Salzberg, S. L. . (2019). Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. *Nature Biotechnology*, 37(8), 1.
- Li, F., G. Fan, K. Wang, F. Sun, Y. Yuan, G. Song, Q. Li, Z. Ma, C. Lu, C. Zou, W. Chen, X. Liang, H. Shang, W. Liu, C. Shi, G. Xiao, C. Gou, W. Ye, X. Xu, X. Zhang, H. Wei, Z. Li, G. Zhang, J. Wang, K. Liu, R. J. Kohel, R. G. Percy, J. Z. Yu, Y. X. Zhu, J. Wang and S. Yu (2014). "Genome sequence of the cultivated cotton *Gossypium arboreum*." Nat Genet **46**(6): 567-572.
- Lightfoot, D. J., D. E. Jarvis, T. Ramaraj, R. Lee, E. N. Jellen and P. J. Maughan (2017). "Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (*Amaranthus hypochondriacus*) chromosomes provide insights into genome evolution." BMC Biol **15**(1): 74.
- Paterson, A. H., J. F. Wendel, H. Gundlach, H. Guo, J. Jenkins, D. Jin, D. Llewellyn, K. C. Showmaker, S. Shu, J. Udall, M. J. Yoo, R. Byers, W. Chen, A. Doron-Faigenboim, M. V. Duke, L. Gong, J. Grimwood, C. Grover, K. Grupp, G. Hu, T. H. Lee, J. Li, L. Lin, T. Liu, B. S. Marler, J. T. Page, A. W. Roberts, E. Romanel, W. S. Sanders, E. Szadkowski, X. Tan, H. Tang, C. Xu, J. Wang, Z. Wang, D. Zhang, L. Zhang, H. Ashrafi, F. Bedon, J. E. Bowers, C. L. Brubaker, P. W. Chee, S. Das, A. R. Gingle, C. H. Haigler, D. Harker, L. V. Hoffmann, R. Hovav, D. C. Jones, C. Lemke, S. Mansoor, M. ur Rahman, L. N. Rainville, A. Rambani, U. K. Reddy, J. K. Rong, Y. Saranga, B. E. Scheffler, J. A. Scheffler, D. M. Stelly, B. A. Triplett, A. Van Deynze, M. F. Vaslin, V. N. Waghmare, S. A. Walford, R. J. Wright, E. A. Zaki, T. Zhang, E. S. Dennis, K. F. Mayer, D. G. Peterson, D. S. Rokhsar, X. Wang and J. Schmutz (2012). "Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres." Nature **492**(7429): 423-427.

- Pertea, M. , Pertea, G. M. , Antonescu, C. M. , Chang, T. C. , Mendell, J. T. , & Salzberg, S. L. . (2015). Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. *Nature Biotechnology*, 33(3), 290-295.
- Salzberg, S. L., A. M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. J. Treangen, M. C. Schatz, A. L. Delcher, M. Roberts, G. Marcais, M. Pop and J. A. Yorke (2012). "GAGE: A critical evaluation of genome assemblies and assembly algorithms." *Genome Res* **22**(3): 557-567.
- Servant, N., N. Varoquaux, B. R. Lajoie, E. Viara, C. J. Chen, J. P. Vert, E. Heard, J. Dekker and E. Barillot (2015). "HiC-Pro: an optimized and flexible pipeline for Hi-C data processing." *Genome Biol* **16**: 259.
- Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov(2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31: 3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>
- Smit, A. F. A., R. Hubley, and P. Green (2019). RepeatMasker Open-4.0. <http://www.repeatmasker.org>
- van Berkum, N. L., E. Lieberman-Aiden, L. Williams, M. Imakaev, A. Gnirke, L. A. Mirny, J. Dekker and E. S. Lander (2010). "Hi-C: a method to study the three-dimensional architecture of genomes." *J Vis Exp*(39).
- Wang, K., Z. Wang, F. Li, W. Ye, J. Wang, G. Song, Z. Yue, L. Cong, H. Shang, S. Zhu, C. Zou, Q. Li, Y. Yuan, C. Lu, H. Wei, C. Gou, Z. Zheng, Y. Yin, X. Zhang, K. Liu, B. Wang, C. Song, N. Shi, R. J. Kohel, R. G. Percy, J. Z. Yu, Y.-X. Zhu, J. Wang and S. Yu (2012). "The draft genome of a diploid cotton *Gossypium raimondii*." *nature genetics* **44**: 7.
- Xie, T., J. F. Zheng, S. Liu, C. Peng, Y. M. Zhou, Q. Y. Yang and H. Y. Zhang (2015). "*De novo* plant genome assembly based on chromatin interactions: a case study of *Arabidopsis thaliana*." *Mol Plant* **8**(3): 489-492.
- Yuan, D., Z. Tang, M. Wang, W. Gao, L. Tu, X. Jin, L. Chen, Y. He, L. Zhang, L. Zhu, Y. Li, Q. Liang, Z. Lin, X. Yang, N. Liu, S. Jin, Y. Lei, Y. Ding, G. Li, X. Ruan, Y. Ruan and X. Zhang (2015). "The genome sequence of Sea-Island cotton (*Gossypium barbadense*) provides insights into the allopolyploidization and development of superior spinnable fibres." *Sci Rep* **5**: 17662.
- Zhang, J., G.-Q. Huang, D. Zou, J.-Q. Yan, Y. Li, S. Hu and X.-B. Li (2018). "The cotton (*Gossypium hirsutum*) NAC transcription factor (FSN1) as a positive regulator participates in controlling secondary cell wall biosynthesis and modification of fibers." *New Phytologist* **217**(2): 625-640.

Tables

Table 1 The statistics of pseudo-chromosome length.

We named the reassemble chromosome by its length order. The length of chromosome contained the 100 bp N between neighboring scaffolds.

Pseudo-chromosome	Length (bp)
chr1	53,706,613
chr2	58,208,992
chr3	49,533,501
chr4	60,231,643
chr5	63,950,048
chr6	49,383,601
chr7	65,258,068
chr8	56,972,542
chr9	68,312,541
chr10	60,016,945
chr11	62,276,721
chr12	37,762,904
chr13	59,096,274
Total anchored	744,710,393
Unanchored	14,054,600

Table 2 The statistics of the draft genome.

All of the statistics got rid of the short scaffolds (scaffold length < 1000bp).

Parameters	Contigs (bp)	length (bp)	Scaffolds (bp)	length (bp)	Contigs number	Scaffolds number
Total	716,234,346		756,905,237		37,849	2,582
Max_length	333,622		10,920,000		-	-
Number>=2000bp	-		-		27,412	1,626
N50	44,885		1,600,000		4,810	139
N60	35,741		1,216,543		6,597	194
	27,678		947,408		8,872	264

Table 3 The statistics of the reassembled *G. raimondii* genome.

All of the statistics got rid of the short scaffolds (scaffold length < 1000bp).

Parameters	Contigs (bp)	length (bp)	Scaffolds (bp)	length (bp)	Contigs number	Scaffolds number
Total	716,283,110		757,086,669		37,921	1,329
Max_length	333,622		73,966,406		-	-
Number>=2000bp	-		-		27,412	462
N50	44,885		60,016,944		4,810	6
N60	35,740		58,208,991		6,598	8

Table 4: The statistics of clustering results.	
Sample	Gossypium raimondii
Number of Sequence in Draft Genome	4,974
Length of Sequence in Draft Genome /bp	758,633,485
Number of Sequence in Clustering	2,883
Rate of Numbers in Clustering /%	57.96
Length of Sequence in Clustering /bp	746,659,745
Rate of Length in Clustering /%	98.42

Table 5: The statistics of ordering results.	
Sample	Gossypium raimondii
Number of Sequence in Ordering	1,328
Rate of Number in Ordering /%	46.06
Length of Sequence in Ordering	744,578,885
Rate of Length in Ordering /%	99.72
Number of Sequence in Trunks	697
Rate of Number in Trunks /%	52.48
Length of Sequence in Trunks	724,960,878
Rate of Length in Trunks /%	97.37

Table 6: The statistics of orienting results.	
Sample	Gossypium raimondii
Number of Sequence in Orienting	1,129
Rate of Numbers in Orienting /%	85.02
Length of Sequence in Orienting /bp	743,948,690
Rate of Length in Orienting /%	99.92

Table 7: The statistics of pseudo-chromosome length.		
Pseudo-chromosome	Scaffolds number	Length /bp
chr1	171	73,966,406
chr2	98	65,258,067
chr3	93	63,950,047
chr4	94	62,276,720
chr5	117	60,231,642
chr6	76	60,016,944
chr7	88	59,096,273
chr8	107	58,208,991
chr9	69	56,972,541
chr10	125	53,706,612
chr11	90	51,080,163
chr12	99	49,533,500
chr13	101	30,412,479
Total anchored	1,328	744,710,385
Unanchored	3,646	14,054,600
We named the reassemble chromosome by its length order. The length of chromosome contained the 100 bp N between neighboring scaffolds.		

Table 8: The statistics of the draft genome.				
Parameters	Contigs length /bp	Scaffolds length /bp	Contigs number	Scaffolds number
Total	716,234,346	756,905,237	37,849	2,582
Max_length	333,622	10,920,000	-	-
Number>=2000bp	-	-	27,412	1,626
N50	44,885	1,600,000	4,810	139
N60	35,741	1,216,543	6,597	194
N70	27,678	947,408	8,872	264
N80	19,881	721,270	11,905	355
N90	11,331	439,650	16,558	485
All of the statistics got rid of the short scaffolds (scaffold length < 1,000bp).				

Table 9: The statistics of the reassembled <i>Gossypium raimondii</i> genome.				
Parameters	Contigs length /bp	Scaffolds length /bp	Contigs number	Scaffolds number
Total	716,283,110	757,086,669	37,921	1,329
Max_length	333,622	73,966,406	-	-
Number>=2000bp	-	-	27,412	462
N50	44,885	60,016,944	4,810	6
N60	35,740	58,208,991	6,598	8
N70	27,675	56,972,541	8,874	9
N80	19,873	53,706,612	11,907	10
N90	11,327	49,533,500	16,562	12

All of the statistics got rid of the short scaffolds (scaffold length < 1,000bp).

Supplemental File 2

Supplemental 2: The reassemble genome.

Supplemental file 2 was not provided with this version of the manuscript.

Figures

G_raitmondii resolution=100000
Genome-wide all-by-all Hi-C interaction

Figure 1

The heat-map of the reassembled result at a 100-kb resolution. Hi-C interactions detected in the reassembled *G. raimondii* genome sequences, Most interactions maps of chromosome sequences suggested that the genome sequence was assembled in the correct order. The darker red dot is representing the higher interaction between bins.

Figure 2

Genome comparisons between the draft genome and reassembled genome with two previous genome assemblies of *G. raimondii* () A) Genome alignment between *G. raimondii* (Wang et al. 2012) and *G. raimondii* (Paterson et al. 2012). B) Genome alignment between *G. raimondii* (Wang et al. 2012) and *G. raimondii* (Udall et al. 2019). C) Genome alignment between *G. raimondii* (new) and *G. raimondii* (Paterson et al. 2012). D) Genome alignment between *G. raimondii* (new) and *G. raimondii* (Udall et al. 2019).

Figure 3

The collinearity analysis of homologous genes between the reassemble genome, the draft genome (Wang et al.2012) and reassembled genome (new) with two previous genome assemblies of *G. raimondii* (Paterson et al. 2012, Udall et al. 2019).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [errorbin.bed.xls](#)