Contents lists available at ScienceDirect

Industrial Crops & Products

journal homepage: www.elsevier.com/locate/indcrop

Comparative analysis of genome sequences of the two cultivated tetraploid cottons, *Gossypium hirsutum* (L.) and *G. barbadense* (L.)

Qingying Meng[a,b], Jiaqi Gu[a,b], Zhongping Xu[a,b], Jie Zhang[a,b], Jiwei Tang[a,b], Anzhou Wang[b], Ping Wang[a,b], Zhaowei Liu[a,b], Yuxuan Rong[a,b], Peihao Xie[a,b], Liuyang Hui[a,b], Joshua A. Udall[c], Corrinne E. Grover[d], Jonathan F. Wendel[d], Shuangxia Jin[a,b], Xianlong Zhang[a,b], Daojun Yuan[a,b,*]

[a] *National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, Hubei 430070, China*
[b] *College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, Hubei 430070, China*
[c] *USDA/Agricultural Research Service, Crop Germplasm Research Unit, College Station, TX 77845, USA*
[d] *Department of Ecology, Evolution, and Organismal Biology (EEOB), Bessey Hall, Iowa State University, Ames, IA 50011, USA*

ARTICLE INFO

Keywords:
Cotton genome
Genome assembly quality
Comparative genomic
Gene index
Centromeric region

ABSTRACT

With innovations in sequencing technology and the progress of high-performance computing systems, it is now relatively straightforward to sequence and assemble complex genomes. Many genomes from multiple cotton species have been released in recent years, with the highly homozygous standard genetic lines of two cultivated allotetraploid cottons, *i.e., Gossypium hirsutum* TM-1 and *G. barbadense* 3–79, assembled multiple times by different research groups using diverse sequencing technologies. The assembly quality among these genomes is variable, even between multiple accessions or versions of the same species, which can generate both confusion in choosing the appropriate genome for genetic analysis and obstacles when comparing results among the different reference genomes. Accordingly, an assessment of the many cotton genome sequences is necessary to facilitate both choice of genome sequence and comparisons between different versions or species. Here we comprehensively assess and compare genome assembly accuracy, completeness, and contiguity for nine *G. hirsutum* assemblies and four *G. barbadense* assemblies using multiple analysis strategies with the same criteria. We identify centromeric regions and several large-scale inversions among genomes from the same accession, indicating structural errors introduced during sequence ordering and orientation in *G. hirsutum* and *G. barbadense* genome assembly. Gene relationships between annotations from multiple genomes are defined within and across species, and the results are available at the Cotton Paralogs Groups Search website (https://ihope.shinyapps.io/cotton-Paralogs/), a convenient resource for converting gene ids and comparing annotations between different genome versions. This study comprehensively assesses and compares assembly quality among multiple versions of the two cultivated tetraploid cotton species with different assembly strategies, illustrating the challenges of sequencing and assembling complex genomes and providing a resource for cotton genomics.

1. Introduction

Cotton (*Gossypium* L.) is a globally important fiber and oil seed crop, and as such, it has been the subject of considerable scientific interest (Wendel and Grover, 2015). The process of cotton fiber initiation and elongation represents one of the best models for deciphering mechanisms of single-cell differentiation and growth (Wang et al., 2020). In addition, tetraploid cotton is an attractive model for studying the origin, evolution and domestication of polyploid species (Hu et al., 2019). While the cotton genus is best known for its crop species, it exhibits extraordinary diversification among the approximately 45 diploid species comprising the eight diploid genome groups (designated A-G, and K) and 7 allotetraploid species (AD) (Endrizzi et al., 1985; Gallagher et al., 2017; Grover et al., 2015). Allotetraploid cotton originated from a hybridization and chromosome doubling event approximately 1–2 million years ago (mya), that occurred in the New World and involved

[*] Corresponding author at: National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, Hubei 430070, China. Fax: +86-027-87280196

E-mail address: robert@mail.hzau.edu.cn (D. Yuan).

https://doi.org/10.1016/j.indcrop.2023.116471
Received 28 September 2022; Received in revised form 23 December 2022; Accepted 17 February 2023
Available online 1 March 2023

an African A genome-like species and a native D genome-like species (Wendel, 1989). Notably, four cotton species were independently domesticated in diverse geographic regions, including two A-genome diploid cotton species (*G. herbaceum* and *G. arboreum*) on the Indian subcontinent and two allotetraploid cotton species (*G. hirsutum* and *G. barbadense*) in Central and northern South America, respectively. While all four species are still grown, the high-yielding crop species *G. hirsutum* has been adapted for and adopted in most growing regions that traditionally cultivated diploid cotton, while the moderately-producing *G. barbadense* is grown at a lower market share for its superior fiber quality that confers a competitive advantage to specialty cotton textiles.

Decoding cotton genomics is the foundation for deciphering cotton evolution, detecting agronomic traits, understanding gene function, and improving cotton breeding, among other things. The polyploid nature of the dominant cultivated species, however, makes genome assembly challenging. To overcome the technical obstacles associated with polyploidy, a coalition of cotton genome scientists previously developed a strategy for sequencing cotton genomes, targeted toward expand opportunities for cotton research and improvement worldwide (Chen et al., 2007). This strategy involved first sequencing the two model diploid progenitors of allopolyploid cotton to establish a foundation for sequencing the cultivated polyploid species. *G. raimondii*, the closest living relative to the paternal ancestor of the allopolyploid D subgenome, was targeted first, both due to its relationship to the cultivated allotetraploids and its small genome size (880 Mb; (Hendrix and Stewart, 2005)), thereby providing fundamental information regarding gene content and organization. An initial sequencing effort was completed using whole-genome shotgun (WGS) sequencing in 2012, producing 775.2 Mb in scaffolded sequence, of which approximately 73.2 % was anchored and only 52.4 % was both anchored and oriented (Wang et al., 2012a). This was shortly followed by a separately generated "gold-standard" *G. raimondii* assembly that integrated ABI 3730XL capillary, Roche 454 XLR, and Illumina GAIIx-derived sequencing (Paterson et al., 2012). Subsequently, *de novo* genome sequences of additional D-genome diploid cotton species have been assembled using long-read sequencing technologies, including a third, independently generated version of the *G. raimondii* genome, which found structural errors in the "gold standard" version (Udall et al., 2019; Wang et al., 2021). Shortly after the release of the initial *G. raimondii* genomes, a draft genome from the model maternal allopolyploid progenitor, *G. arboreum*, was released (Li et al., 2014), which was later supplemented by three additional independent assemblies (Du et al., 2018; Huang et al., 2020; Wang et al., 2021). The sister species to *G. arboreum* (and equivalent model maternal progenitor), *i.e., G. herbaceum*, was later sequenced by two different groups (Huang et al., 2020; Ramaraj et al., 2022), bringing the total number of genome sequences for the allopolyploid progenitors up to seven. As with *G. raimondii*, later sequencing efforts leveraged long-read sequencing technologies, leading to improvements in genome contiguity; however, differences still remain among similarly generated sequences from the same species. While not the focus of the present paper, the cotton genus as a whole has recently experienced a flurry of long-read based genome assemblies from diploids of increasing phylogenetic distance from the two model diploid progenitors (Cai et al., 2020; Grover et al., 2020, 2021a, 2021b; Perkin et al., 2021; Sheng et al., 2022; Udall et al., 2019; Xu et al., 2022), further increasing genome availability.

Shortly after the initial diploid genome sequences were generated, attention turned to tackling the complex genomes of the highly valuable tetraploid cultivated species. As with the diploids, multiple accessions from both *G. hirsutum* and *G. barbadense* were independently sequenced numerous times, first relying primarily on Illumina sequencing but later transitioning to long-read sequencing, leading to the high-quality tetraploid genomes available today. The genetic standard *G. hirsutum* accession TM-1 has been sequenced eight times (Chen et al., 2020; Hu et al., 2019; Huang et al., 2020; Li et al., 2015; Saski et al., 2017; Wang et al., 2019; Yang et al., 2019b; Zhang et al., 2015), while the *G. barbadense* accession '3–79′ has been sequenced thrice (Chen et al., 2020; Wang et al., 2019; Yuan et al., 2015). Additional polyploid genomes are also available for the *G. hirsutum* cv. ZM24 (Yang et al., 2019b), cv. XLZ7 (He et al., 2021), cv. Bar32–30 (Perkin et al., 2021), cv. Barbren-713(Perkin et al., 2021) and cv. NDM8 (Ma et al., 2021); for *G. barbadense* cv. Xinhai21 (Liu et al., 2015), cv. Hai7124 (Hu et al., 2019), and cv. Pima90 (Ma et al., 2021); for five wild tetraploid species, *G. tomentosum* [$(AD)_3$], *G. mustelinum* [$(AD)_4$], *G. darwinii* [$(AD)_5$], *G. ekmanianum* [$(AD)_6$] and *G. stephensii* [$(AD)_7$]; and for the relatively unimproved *G. hirsutum* race punctatum (Chen et al., 2020; Peng et al., 2022).

Previous studies on cotton assembly quality assessment mainly focused on the general assembly statistics and broad collinearity (*via* dotplots or similar), but without uniform criteria across studies. Because the assembly quality of a reference genome could impact the accuracy of the genetic and/or comparative analyses based upon it, it is critical to choose the highest quality reference genome for a given application. With more than eight *G. hirsutum* TM-1 and three *G. barbadense* 3–79 assemblies available, a comprehensive, multidimensional comparison becomes paramount in choosing the appropriate version, and understanding the differences among assemblies becomes an important factor in comparing observations among studies both present and future.

This study aims to present a detailed comparison of the assemblies available for the two cultivated allotetraploid cotton genomes with multiple sequences available for a single accession, *i.e.*, *G. hirsutum* TM-1 and *G. barbadense* 3–79. This research presents a comprehensive evaluation of assembly contiguity, accuracy, and completeness from each of these genomes, identifies the centromeric regions using CenH3 ChIP-seq data, and establishes synteny and gene orthology relationships between different versions. Our assessment suggests that the quality of genomes assembled with PacBio long reads are much better than those assembled from Illumina short reads. For the PacBio-based assemblies, the quality of TM1_CRI_v1 and 3–79_HAU_v2 is a little higher than other *G. hirsutum* TM-1 and *G. barbadense* 3–79 assemblies, respectively.

2. Materials and methods

2.1. Data collection

Genome sequences and annotations for the two cultivated tetraploid species, *G. hirsutum* TM-1 and *G. barbadense* 3–79 (with the exception of 3–79_UTX_v1.1), and their model diploid ancestors (maternal: *G. arboreum* or *G. herbaceum*; paternal: *G. raimondii*) were downloaded from the publicly available database CottonGen (Yu et al., 2021). The 3–79_UTX_v1.1 genome (Chen et al., 2020) was downloaded from NCBI (GCA_008761655.1). Eight *G. hirsutum* TM-1 assemblies are available, including TM1_BGI_v1 (Li et al., 2015), TM1_NBI_v1.1 (Zhang et al., 2015), TM1_JGI_v1.1 (Saski et al., 2017), TM1_HAU_v1 (Wang et al., 2019), TM1_ZJU_v2.1 (Hu et al., 2019), TM1_CRI_v1 (Yang et al., 2019b), TM1_UTX_v2.1 (Chen et al., 2020), TM1_WHU_v1 (Huang et al., 2020). The *G. barbadense* assemblies analyzed here were 3–79_HAU_v1 (Yuan et al., 2015), 3–79_HAU_v2 (Wang et al., 2019), and 3–79_UTX_v1.1 (Chen et al., 2020). Two assemblies ZM24_CRI_v1 (Yang et al., 2019b) and Hai7124_ZJU_v1.1 (Hu et al., 2019) are also available. Additionally, genome sequences for the model diploid progenitors were downloaded, which include Mutema_A1_WHU_v1 (Huang et al., 2020), SXY1_WHU_v1 (Huang et al., 2020), and D5_NSF_v1 (Udall et al., 2019).

2.2. BAC sequence alignment

We downloaded 193 BAC sequences that are reportedly from *G. hirsutum* in NCBI (Guo et al., 2008; Hu et al., 2019; Wang et al., 2019; Yang et al., 2019a). These BAC sequences were aligned to the nine *G. hirsutum* genomes by MUMmer v3.23 (Delcher et al., 2003) using the following procedure. First, all 193 sequences were aligned to genome using the nucmer utility with the parameter `-mum`. Then, to filter mapping noise and determine the one-to-one alignment blocks, the parameter with `– 1 -i 90 -l 1000` was used, resulting in 164 BACs retained. Finally, coordinates of each BAC sequence on the chromosomes was recovered using show-coords parameter `-rcl -TH`.

2.3. BUSCO evaluation

The assembly completeness of *G. hirsutum* and *G. barbadense* with respect to gene content was evaluated using BUSCO v4.0.4 [27] with the "embryophyte_odb10″ dataset under default parameters. A total of 1614 highly conserved genes were used to search the genome assembly, the annotated proteins, and the transcriptome. We also evaluated the genome, transcriptome, and protein completeness of the three putative diploid ancestor genomes using the same BUSCO-based method.

2.4. EST and GSS sequences alignment

Expressed sequence tags (ESTs) and genome survey sequences (GSS), representing the expressed and genomic portions of the genome respectively, were downloaded from NCBI for each species. All sequences were filtered for vector contamination using vecscreen_plus_taxonomy v2.0.0 (Schäffer et al., 2017) with default parameters. Subsequently, sequence redundancy was reduced by CD-HIT v4.8.1 (Li and Godzik, 2006) with parameters `-c 0.95`. This processing resulted in 132,995 *G. hirsutum* ESTs (average length = 733 bp); 16,886 *G. barbadense* ESTs (average length = 675 bp); and 180,521 *G. hirsutum* GSSs (average length = 639 bp). These sequences were aligned to their corresponding genome sequences using BLAT (Kent, 2002) with default parameters. A strict cutoff of 90 % coverage and 90 % identity was set for the GSS sequences, and a relatively loose cutoff of 50 % coverage and 90 % identity was set for the EST sequences.

2.5. PacBio transcriptome reads mapping

G. hirsutum and *G. barbadense* full-length RNA-seq sequenced *via* PacBio technology was downloaded from NCBI (BioProject: PRJNA433615, PRJNA493958, PRJNA503814, PRJNA503326 and PRJNA359724). Then, reads with more than 100 bp were mapped to their genome assemblies using minimap2 v2.17 (Li, 2018) with parameters "-ax splice:hq". Subsequently, the mapping ratio for each genome was calculated using the stats utility from SAMtools v1.9 (Li et al., 2009) with default parameters.

2.6. Assessment of the assembly contiguity with LAI

LTR-RT candidates were obtained using (1) LTRharvest v1.6.1 (Ellinghaus et al., 2008) with parameters '-minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1 -similar 85 -vic 10 -seed 20 -seqids yes' and (2) LTR_FINDER_parallel v1.1 (Ou and Jiang, 2019) with parameters '-harvest_out -size 1000000 -time 300′. After identifying LTR-RTs and generating high-quality LTR libraries with LTR_retriever v2.9.0 (Ou and Jiang, 2018), the LTR Assembly Index (LAI) for each assembly was calculated with default parameters (Ou et al., 2018).

2.7. Identification of the centromeric regions by CenH3 ChIP

ChIP-seq data were downloaded from NCBI (PRJNA488416). After filtering low-quality reads using Trimmomatic v0.43 (Bolger et al., 2014) with the parameters 'LEADING:10 SLIDINGWINDOW:4:15 MINLEN:50′, FastUniq (Xu et al., 2012) was used to remove duplicated read pairs under default parameters. High-quality Illumina reads were mapped to their genome assemblies with Bowtie2 v2.3.2 (Langmead and Salzberg, 2012) and the parameter `-N 1`, and the output was converted to sorted Binary Alignment Map (BAM) using SAMtools v1.9 (Li et al., 2009). Only high-quality mapping reads (-F 4 -q 30) were kept for the further analysis. The number of mapped reads was counted for each 10 kb non-overlapping window. The read density was calculated by dividing the total number of mapped reads by the total number of mapped nucleotides in each genomic window. In order to remove the impact of non-specific binding by rabbit serum, the read density was adjusted for background signal by using a mock control data. CenH3 domains were identified *via* SICER2 with ` -g 400 –significant_reads` (Zang et al., 2009). The CenH3 domains were defined as those where the fold-change/control was ≥ 5 and the false-discovery rate (FDR) was < 0.01, using 200-bp windows and allowing gaps of 400 bp.

2.8. Synteny block identification

Syntenic blocks were identified by whole genome alignment with MUMmer v3.2.3 (Delcher et al., 2003). Alignment of the genomes was performed using the nucmer utility under the parameters `-mum`, and then the alignment block was filtered using delta-filter with one-to-one alignment mode. Alignment coordinates were obtained with show-coords. Rearrangements and local sequence differences between the genome from the same species were identified *via* SyRI v1.6 (Goel et al., 2019). Then JCVI (https://github.com/jcvi) and dotPlotly (https://github.com/tpoorten/dotPlotly) were used to visualize the alignment results.

2.9. One-to-one gene relationship identified

In order to identify syntenic genes, protein sequences were compared by all-*versus*-all BLASTP v2.10.0 with the parameters `-evalue 10e-5 -outfmt 6 -num_alignments 15` (Schäffer et al., 2001). Then, the putatively homologous genes identified by BLASTP were analyzed for synteny by the MCScanX package using default settings (Wang et al., 2012b). Syntenic blocks were defined as those with at least five syntenic genes.

2.10. Hi-C sequence data processing and visualization

The Hi-C sequences for TM1_UTX_v2.1 and 3–79_UTX_v1.1 are unavailable, the Hi-C sequences for the *G. hirsutum* TM1 (Huang et al., 2020), *G. barbadense* 3–79 (Wang et al., 2018) and *G. barbadense* Hai7124 (Hu et al., 2019) were download from NCBI (BioProject: PRJNA524970 for TM1, PRJNA396502 for 3–79, for PRJNA505106 for Hai7124). The HiC-Pro pipeline (v2.11.14) was used to generate chromatin interaction matrices (Servant et al., 2015). Briefly, Hi-C reads were filtered by Trimmomatic v0.43 (Bolger et al., 2014) and then mapped to the corresponding assemblies using bowtie2 (Langmead and Salzberg, 2012). An iterative correction and eigenvector decomposition (ICE) method was used to normalize the Hi-C contact matrices. Normalized contact matrices were visualized with juicebox (Durand et al., 2016).

3. Results and discussion

3.1. Basic assembly quality information statistics

A total of thirteen assembles from two cultivated tetraploid cotton genomes were collected, including nine *G. hirsutum* assemblies and four *G. barbadense* (Table 1). Three of these assemblies were generated using Illumina short-read sequence technology, including TM1_BGI_v1, TM1_NBI_v1.1 and 3–79_HAU_v1, while two were generated using the 10X Genomics technology to link newly generated Illumina short-read based assemblies (TM1_ZJU_v2.1 and Hai7124_ZJU_v1.1). The remaining eight assemblies (TM1_JGI_v1.1, TM1_HAU_v1, TM1_CRI_v1, ZM24_CRI_v1, TM1_UTX_v2.1, TM1_WHU_v1, 3–79_HAU_v2, and 3–79_UTX_v1.1) were sequenced by PacBio long-read technology. Some of these were anchored and oriented by combining Hi-C and BioNano

Table 1
Basic genome quality information of two cultivated tetraploid cottons.

Version	TM1_BGI_v1	TM1_NBI_v1.1	TM1_JGI_v1.1	TM1_HAU_v1	TM1_ZJU_v2.1	TM1_CRI_v1	ZM24_CRI_v1	TM1_UTX_v2.1	TM1_WHU_v1	3–79_HAU_v1	3–79_HAU_v2	Hai7124_ZJU_v1.1	3–79_UTX_v2.1
Sequencing Technology	Illumina	Illumina	Pacbio	Pacbio+Hi-C +Bionano	10X Genomic+ Hi-C +Bionano	Pacbio+Hi-C	Pacbio+Hi-C	Pacbio+Hi-C	Pacbio+Hi-C	Illumina	Pacbio+Hi-C +Bionano	10X Genomic+Hi-C +Bionano	Pacbio+Hi-C
Total assembled size (Mb)	2151	2546	2342	2347	2298	2289	2309	2305	2290	2573	2267	2227	2196
Anchored and oriented size (Mb)	1792	1935	2171	2233	2227	2241	2150	2282	2271	1997	2133	2184	2070
Percentage of ordered and oriented (%)	83.31	76.00	92.70	95.14	96.91	97.90	93.11	99.00	99.17	77.61	94.09	98.07	94.26
Gap length (Mb)	61	383	82	65	31	<1	<1	3	<1	335	44	34	2
Contigs Number	42,057	444,325	13,583	4791	51,762	1283	3717	6733	1235	285,315	4943	75,898	4767
Largest contig (kb)	1062	951	4956	25,006	2103	23,174	14,261	9045	27,970	252	19,605	1536	10,615
Contigs N50 (kb)	78	32	389	1892	113	4761	1976	784	5021	25	2152	78	1770
Contigs L50	7897	19,084	1514	280	4807	141	4807	789	128	25,469	270	6855	360
Scaffolds Number	9154	288299	5355	2190	12,098	—	—	1025	—	17,460	3032	11,350	2048
Scaffolds N50 (Mb)	66	66	90	98	108	—	—	108	—	66.7	92.9	101.6	93.8
Scaffolds L50	12	15	11	11	10	—	—	10	—	15	11	10	30
GC contents (%)	34.45	34.11	34.23	34.37	34.72	34.38	34.41	34.36	34.37	33.86	34.20	34.29	34.12
Annotation Gene Number	76,943	70,478	87,800	70,199	72,761	73,624	73,707	74,376	74,350	80,876	71,297	75,071	74,561
BUSCOs (%)	97.34	98.95	99.19	98.88	99.19	99.26	99.19	99.13	99.26	82.28	99.13	99.19	98.64
LAI	5.45	6.23	11.16	12.75	10.52	14.63	14.59	12.96	14.8	3.75	11.96	10.53	13.36

optical physical mapping (*i.e.,* TM1_HAU_v1, 3–79_HAU_v2, TM1_ZJU_v2.1 and Hai7124_ZJU_v1.1; Table 1) and most of the remainder were anchored using Hi-C alone (ZM24_CRI_v1, TM1_CRI_v1, TM1_UTX_v2.1, TM1_WHU_v1, 3–79_UTX_v1.1).

Basic summary statistics were calculated using in-house Perl scripts (qymeng1996/Script: Bioinfomation Script (github.com)). The total assembled length (with gaps) range from 2.15 Gb to 2.57 Gb, similar to the 2.4–2.5 Gb genome sizes for *G. hirsutum* and *G. barbadense* (Table 1) (Hendrix and Stewart, 2005). Completeness of the genome assemblies was reflected in their anchored and oriented sizes and gap lengths. In the three purely Illumina-based short-read assemblies, fewer than 85 % of scaffolds could be assigned to 26 *pseudo*-chromosomes (*i.e.,* TM1_BGI_v1, TM1_NBI_v1 and 3–79_HAU_v1) and there were numerous gaps (*e.g.,* 383 Mb gaps in TM1_NBI_v1.1; Table 1). In contrast, most of the assemblies generated using long read sequencing technologies could be anchored and oriented to place more than 94 % of scaffolds onto 26 *pseudo*-chromosomes while also containing fewer/shorter gaps (*i.e.*, gap length <1 Mb in several of these genomes; Table 1).

Contiguity of each genome sequence was assessed using both the Contig N50 and Contig L50, where the Contig N50 represents the length of the smallest size-ordered contig that brings the sum of the contigs to > 50 % of the total genome assembly size and Contig L50 represents the number of that of contig (size-ordered) that brings the summed length of contigs > 50 % of the genome size. The three Illumina-only assemblies exhibited the lowest Contig N50 values, while the two assemblies (TM1_ZJU_v2.1 and Hai7124_ZJU_v1.1) that used 10x Genomics in combination with Hi-C and BioNano have a slightly improvement of the N50 value. As expected, the genomes generated using long-read technologies exhibited higher Contig N50 values than other genome assemblies (Table 1 and Fig. 1). By definition, these assemblies with higher Contig N50 values also exhibited lower Contig L50 scores, indicating that fewer contigs were required to represent over half the genome. This suggests that while the two 10x Genomics assemblies (TM1_ZJU_v2.1 and Hai7124_ZJU_v1.1) have a slight improvement in sequence contiguity over the three Illumina-only assemblies, the genomes assembled from long reads experienced significantly improved assembly contiguity. The difference is most stark when comparing the three Illumina-only assemblies with the long-read based assembly with the highest contig N50 (*i.e.,* TM1_WHU_v1), whose largest contig is approximately 200x longer than the Illumina-only assemblies (5021 Kb *versus* 25–78 Kb). We also calculated the NG(x) and LG(x) trends for each genome, which is generally congruent with the results revealed by N50 and L50 (Fig. S1). Of the genomes analyzed, two *G. hirsutum* assemblies (*i.e.*, TM1_WHU_v1 and TM1_CRI_v1) and two *G. barbadense* assemblies (*i.e.*, 3–79_HAU_v2 and 3–79_UTX_v1.1) exhibit better sequence contiguity than the other assemblies.

3.2. Genome assembly accuracy assessment

A total of 193 *G. hirsutum* BAC sequences with an average insert size of 105 kb ranging from 36 to 248 kb were downloaded and aligned to the nine *G. hirsutum* genome sequences to assess assembly accuracy. Of these, only 164 BAC sequences passed our filter parameters (see Materials and Methods), with the remaining 29 BAC sequences identified as contaminants originating from either maize or bacterial genomes (Table S2 and Table S3). From the 164 remaining, three BAC sequences exhibited < 10 % coverage match to any of the genome sequences and were therefore discarded, resulting in 161 BAC sequences that remained for further analysis (Table S1). As expected, the five PacBio-based *G. hirsutum* genome assemblies with greatest contiguity (*i.e.,* TM1_HAU_v1, TM1_CRI_v1, ZM24_CRI_v1, TM1_UTX_v1 and TM1_WHU_v1) also exhibited the best alignments with these BAC sequences, averaging 80.7 % of the BAC sequences aligning to the genome (Fig. 2 and Table S1). Additionally, the two *G. hirsutum* assemblies (*i.e.*, TM1_NBI_v1.1 and TM1_BGI_v1) comprised solely of short Illumina reads exhibited the lowest coverage scores (Fig. 2 and Table S1). The

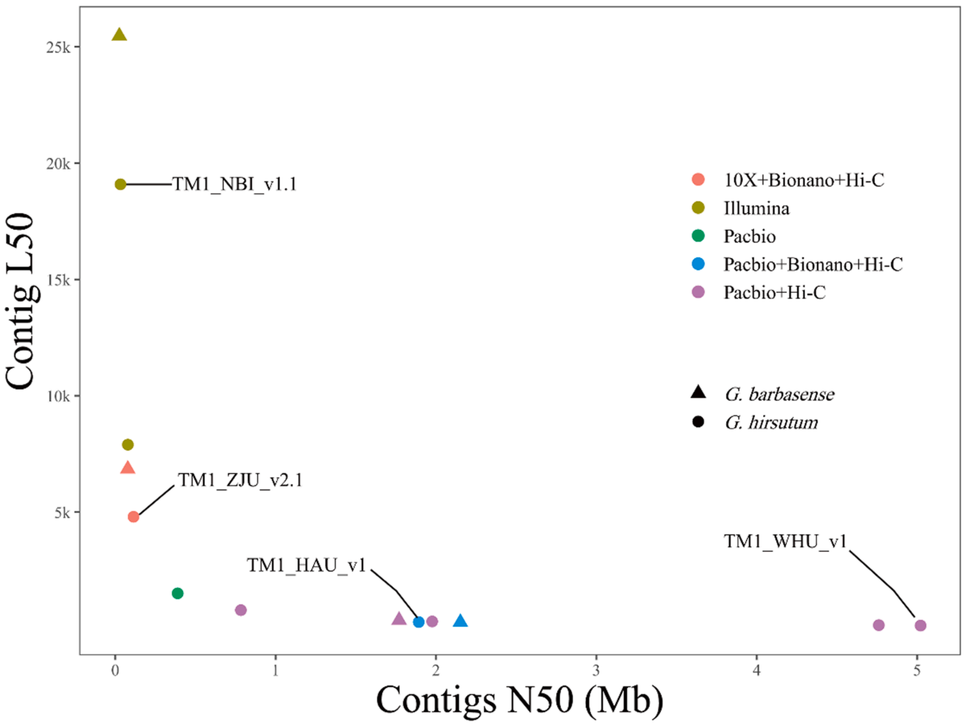

Fig. 1. Contigs L50 and Contigs N50 for the two cultivated tetraploid cotton species genomes, *G. hirsutum* and *G. barbadense*.

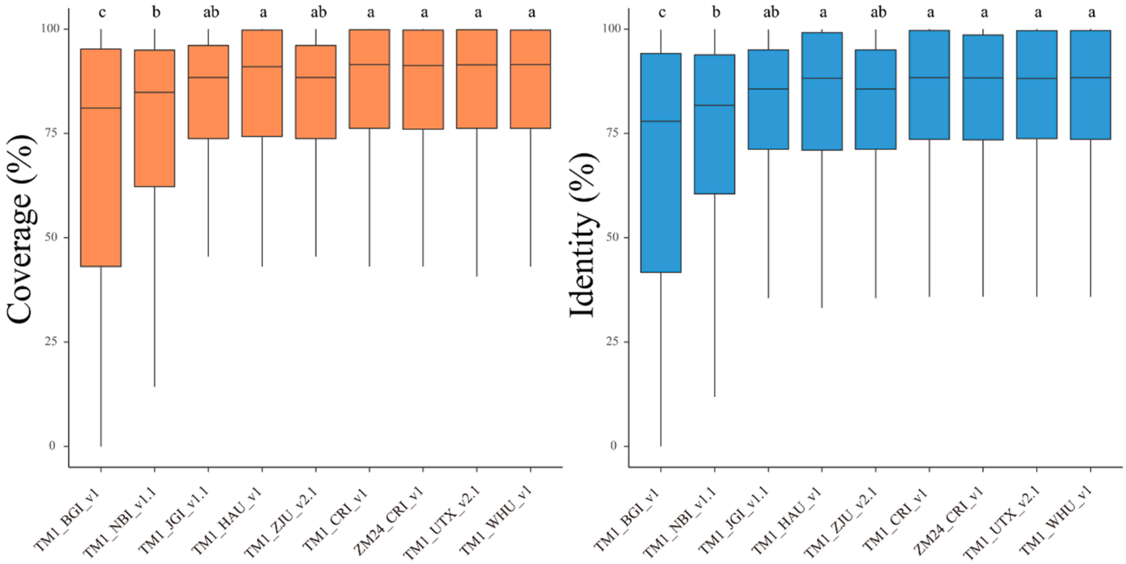

Fig. 2. The Coverage (left) and Identity (right) of *G. hirsutum* BAC sequences align to their genomes with MUMmer (Duncan's test).

trend in alignment scores of these BAC sequences to each genome was similar to the trend in coverage, indicating a general improvement of accuracy and placement when PacBio was used in conjunction with scaffolding technologies such as BioNano and/or Hi-C (designated as PacBio+ hereafter).

We also aligned 180,521 non-redundant *G. hirsutum* genome survey sequences (GSS) to the nine *G. hirsutum* assemblies, with more than 92.0 % of GSS sequences (>500 bp) successfully aligning to each reference genome (Table S4). Notably, the difference among genome versions is minor, with 92.0 % and 93.4 % of GSS sequences aligning to TM1_BGI_v1 and TM1_ZJU_v2.1, respectively, and 94.0–95.2 % aligning to the remaining genomes.

3.3. *Genome assembly completeness assessment*

BUSCO is a widely used software for quantitative assessment of completeness for both genome assembly and annotation based on evolutionarily-informed expectations of gene content. More than 97.3 % complete BUSCOs were identified in both the *G. hirsutum* and *G. barbadense* assemblies (Fig. 3A), with the exception of *G. barbadense* 3–79_HAU_v1, which had 15.4 % missing and 2.3 % fragmented BUSCOs, indicating a potential lack of completeness in the gene space. We found the percentage of duplicated BUSCOs ranged between 64.9 % (TM1_BGI_v1) and 94.6 % (3–79_UTX_v1.1), likely due to the absence of general gene loss post polyploidization (Liu et al., 2001; Zhang et al., 2015) leading to the retention of duplicate BUSCOs derived from each

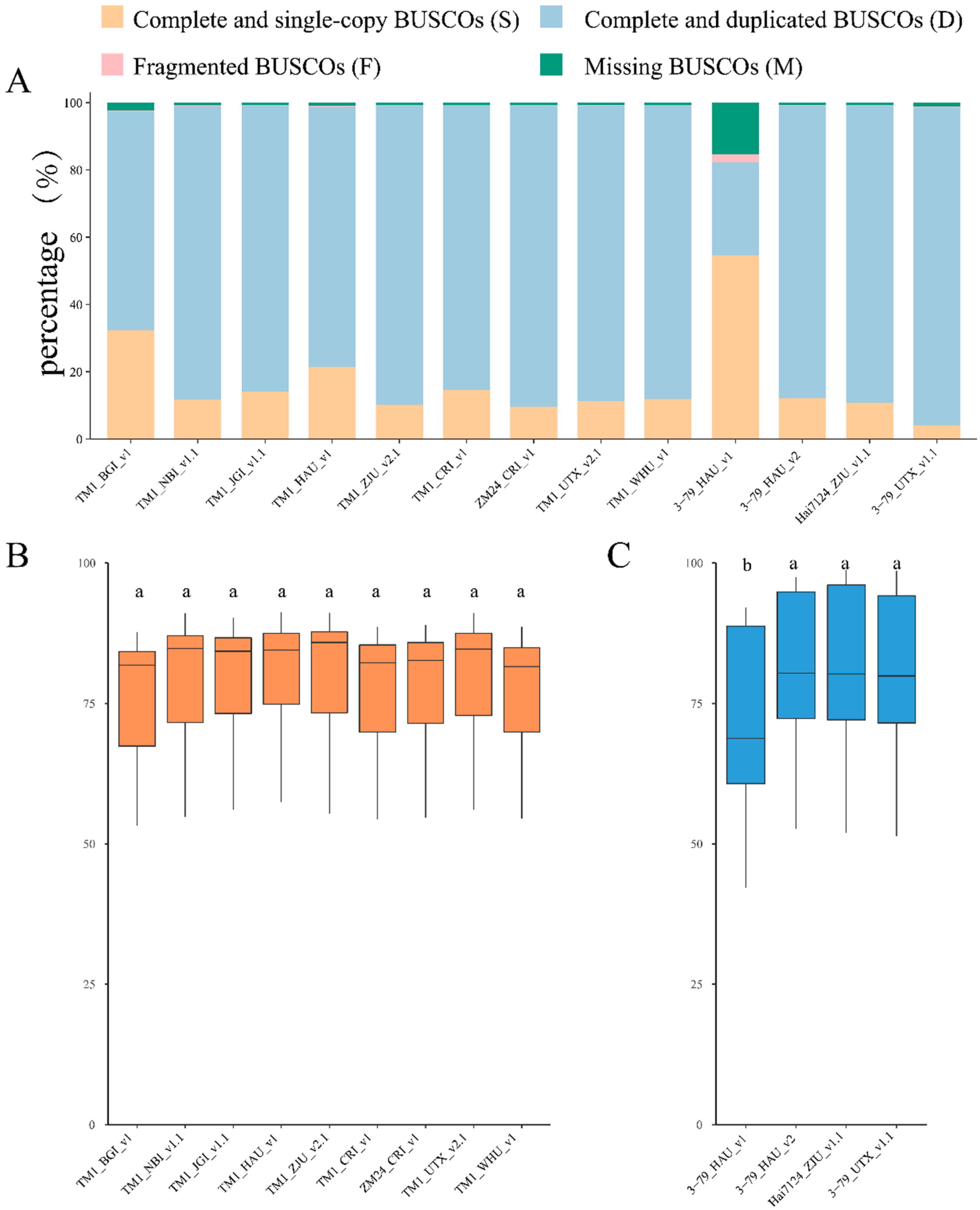

Fig. 3. Genome assembly completement assessment. A BUSCO evaluate the genomes of two tetraploid cotton species. B The percentage of *G. hirsutum* PacBio transcriptome data uniquely mapped to their genomes. C The percentage of *G. barbadense* PacBio transcriptome data uniquely mapped to their genomes.

parent. Therefore, the genomes with a higher proportion of duplicated BUSCOs in these tetraploid cotton genomes are likely more complete. By contrast, more than 94.2 % of complete BUSCOs and 89.2 % of single-copy BUSCOs were identified in the three genome assemblies of the model diploid ancestors of cotton, respectively (Fig. S2C). For transcripts and proteins, we found a similar number of complete BUSCOs in most *G. hirsutum* and *G. barbadense* annotations, with the exception of *G. barbadense* 3–79_HAU_v1, which had far more complete BUSCOs in the transcripts and proteins than in the genome sequence (Fig. S2A and Fig. S2B). Interestingly, the proportion of complete BUSCOs for the both transcriptome and protein predictions was slightly lower than the proportion of BUSCOs recovered from the genome in the diploid genome assemblies (Fig. S2D and Fig. S2E).

The PacBio full-length transcriptomes of *G. hirsutum* and *G. barbadense* were mapped to their genome assemblies to evaluate the gene annotation. The mapping ratio for the PacBio full-length transcriptome in *G. hirsutum* was similar among genomes, ranging from 77.0 % for TM1_BGI_v1 to 80.7 % for TM1_HAU_v1 (Fig. 3B and Table S5). The mapping ratio for the PacBio full-length transcriptome to the three *G. barbadense* assemblies (3–79_HAU_v2, Hai7124_ZJU_v1.1 and 3–79_UTX_v1.1) were also similar, ranging from 80.5 % (3–79_UTX_v1.1) to 81.1 % (Hai7124_ZJU_v1.1, Fig. 3C and Table S6). In both cases, this indicates similar completeness of the gene space.

Expressed sequence tags (ESTs) are single-pass Sanger sequencing reads of approximately 200–800 base pairs (bp) generated from randomly selected cDNA clones, which could provide additional support for confirming the presence of a gene and aid in inferring exon-intron boundaries (Parkinson and Blaxter, 2009). We used available EST sequences to assess genome quality, retaining only ESTs with over 50 % gene-length coverage and over 90 % alignment identity (Tables S7-S8). More than 93.4 % of *G. hirsutum* EST sequences (>500 bp) successfully aligned to each *G. hirsutum* genome, with the exception of TM1_BGI_v1, for which only 92.0 % EST sequences could be aligned. More than 97.4 % of *G. barbadense* EST sequences (>500 bp) successfully aligned to each *G. barbadense* genome, with the exception of 3–79_HAU_v1, for which only 83.1 % EST sequences could be aligned.

3.4. Genome LTR Assembly Index assessment

LTR Assembly Index (LAI) is a reference-free method to assess assembly continuity by evaluating the completeness of a LTR-retrotransposon assembly within a genome (Ou et al., 2018). If the LAI score ranges from 0 to 10, genome sequences are considered as “draft genomes”, but if the range is 10–20, it is considered “reference” grade, and if greater than 20, the genome is considered to be a “Gold” standard. The three Illumina-based short read assemblies had LAI scores less than 10, whereas the LAI scores for the remaining assemblies were greater than 10. The two 10X Genomics (TM1_ZJU_v2.1 and Hai7124_ZJU_v1.1) had higher LAI scores, but not as high as the PacBio assemblies (Fig. 4). In particular, the two *G. hirsutum* assemblies (TM1_WHU_v1 and TM1_CRI_v1) and the two *G. barbadense* assemblies (3–79_UTX_v1.1 and 3–79_HAU_v2) had the highest LAI scores. To visualize the local assembly quality of each assembly, a 3MB-sliding window with a 300-Kb step found that euchromatin regions had higher LAI scores than heterochromatic regions (Figs. S3-S4). Most of regional LAI score for PacBio+ assemblies were larger than 10, and, interestingly, the LAI score for the transposable element enriched A-subgenome (At) is greater than the LAI score for the more compact D-subgenome (Figs. S3-S4).

3.5. Identification of centromeric regions

Centromeric regions are challenging to assemble for *de novo* genome sequences because as they primarily are composed of highly repeated retrotransposon-like and satellite sequences (Han et al., 2016). Genome-wide characterization of DNA sequences associated with CenH3 nucleosomes in *G. hirsutum* and *G. barbadense* (with data such as ChIP-seq) facilitate identification of centromeres in tetraploid cotton. The centromeric regions in *G. hirsutum* TM1_ZJU_v2.1 and *G. barbadense* Hai7124_ZJU_v1.1 were similar to those previously identified (Hu et al., 2019) (Tables S9 and S10), being present on 24 and all 26 chromosomes, respectively. In the Illumina-based short-read assemblies, most *pseudo*-chromosomes are missing centromeric regions, as evidenced by the lack of ChIP-seq alignment (Table 2). In most of 10x Genomics and PacBio-based assemblies, the contiguity and completeness of repetitive sequences were allowable to identify the centromeric regions by ChIP-seq, with the exception of TM1_JGI_v1.1 (PacBio only). Identified centromeric regions in the remaining *G. hirsutum* assemblies occupied 14–25 of the total chromosomes (between 9.40 and 25.41 MB, Table 2), with an average length of 0.94 Mb (Table 2). The centromeric average length of ZM24_CRI_v1 was shortest of all PacBio+ *G. hirsutum* assemblies, while the TM1_HAU_v1 and TM1_ZJU_v2.1 (PacBio+ and 10x Genomics, respectively) both had centromeric regions longer than 1 Mb. At the chromosome level, the D08 centromere could not be identified in any of *G. hirsutum* assemblies, which may be due to *G. hirsutum* D08 containing longer repetitive sequences or the inability of the ChIP-seq to capture the CenH3-bound DNA on this chromosome (Table S9). Similar to the Illumina-based *G. hirsutum* assemblies, the Illumina-based *G. barbadense* assembly (3–79_HAU_v1) quality was lower in repetitive-DNA-rich centromeric regions than other *G. barbadense* assemblies. The three other *G. barbadense* assemblies had identifiable centromeric regions on 21–26 chromosomes, with a total length from 25.02 to 36.57 Mb (Table 2 and Table S10). Only *G. barbadense* Hai7124_ZJU_v1.1 had identifiable centromeric regions from in all 26 assembled chromosomes.

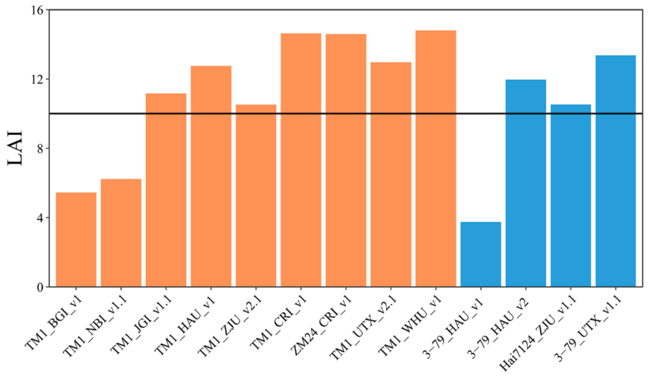

Fig. 4. Genome assembly contiguity assessment. LAI score of *G. hirsutum* (red) and *G. barbadense* (green), a reference level cutoff (LAI = 10) shows by the black horizon line.

3.6. Comparisons of collinearity and gene content within and between the various assemblies of G. hirsutum and G. barbadense

Structural comparisons among the high-quality *G. hirsutum* and *G. barbadense* assemblies can reveal large-scale assembly errors that are not otherwise captured by the preceding analyses. Using whole-genome alignment and focusing on broad-scale synteny, 63 non-redundant inversions and translocations longer than 1 Mb were identified the five TM-1 assemblies (Fig. 5 and Figs. S5-S8). Interestingly, the number of putative rearranged sequences in the A-subgenome (35) was slightly higher than that in the D-subgenome (28) (Table S11), perhaps due to the more repetitive nature of the A-subgenome. Because these differences may reflect segregating polymorphism or technical artifacts, Hi-C heatmaps of TM1 were constructed using existing Hi-C libraries (Huang et al., 2020), which indicated that the large inversions and translocations among the five TM-1 assemblies were likely due to scaffolds being placed in opposite directions in one of the assemblies (Table S11). We found that two genomes (TM1_UTX_v2.1 and TM1_WHU_v1) exhibited 20 and 22 incorrect inversions (*i.e.*, artifacts), respectively, which was several folds greater than those detected in the three other genomes (TM1_HAU_v1, 6 inversions; TM1_CRI_v1, 2 inversions; and TM1_ZJU_v2.1, 7 inversions). In addition to these inversions, 16 translocation assembly artifacts were identified in the TM1_WHU_v1 (Table S11). Overall, three TM-1 genomes, *i.e.,* TM1_HAU_v1, TM1_CRI_v1, and TM1_ZJU_v2.1, had the highest levels of sequence synteny *inter se* and the fewest detected assembly artifacts. Interestingly,

Table 2
Centromeric regions identified by ChIP-seq data.

Version	Centromeric regions	Total length (MB)	Average Length (MB)
TM1_HAU_v1	18	19.40	1.08
TM1_ZJU_v2.1	24	25.41	1.06
TM1_CRI_v1	25	24.64	0.99
ZM24_CRI_v1	14	9.40	0.67
TM1_UTX_v2.1	25	21.48	0.86
TM1_WHU_v1	24	21.92	0.91
3–79_HAU_v1	5	2.71	0.54
3–79_HAU_v2	21	25.34	1.21
Hai7124_ZJU_v1.1	26	36.57	1.41
3–79_UTX_v1.1	25	25.02	1.00

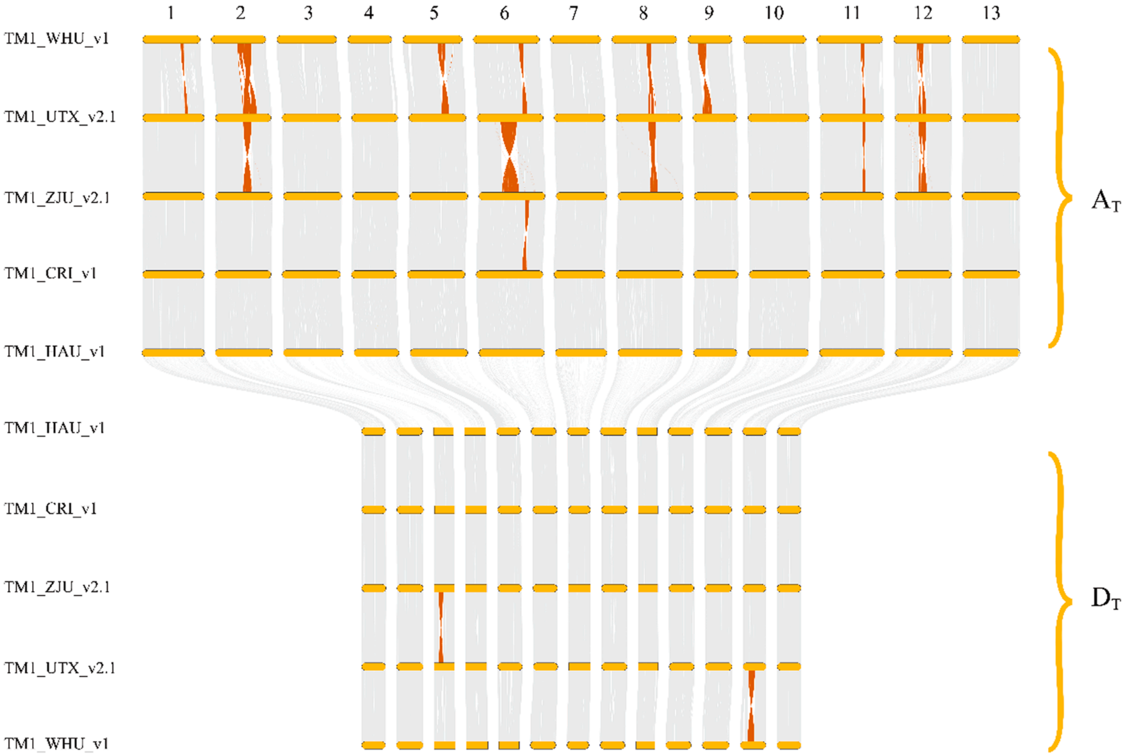

Fig. 5. Genome alignment among five TM-1 reference genomes, the red lines show the rearrangement sequences larger than 5 Mb.

a large inversion on A06 was shared between TM1_UTX_v2.1 and TM1_WHU_v1, despite the Hi-C heatmaps concluding that this inversion is likely artifactually induced during scaffold placement. Similarly, the Hi-C heatmaps also suggest that the larger inversions on A02, A09, and D12 are artifacts of assembly in TM1_UTX_v2.1 or/and TM1_WHU_v1 (Fig. S9).

Syntenic analyses between three *G. barbadense* assemblies, *i.e.,* 3–79_HAU_v2, 3–79_UTX_v1.1 and Hai7124_ZJU_v1.1, were also performed *via* whole-genome alignment (Fig. 6, Figs. S10-S11). These found 36 non-redundant inversions and translocations sequences longer than 1 Mb among the three *G. barbadense* assemblies, most of which were due to incorrect orientation during scaffold placement (Table S12). The number of incorrectly oriented scaffolds varied among the genomes, from only three inversion errors in 3–79_HAU_v2 to 14 and 16 errors in 3–79_UTX_v1.1 and Hai7124_ZJU_v1.1, respectively. Notably, there are 10 large-size, verified inversions between 3 and 79_HAU_v2 and Hai7124_ZJU_v1.1 (*e.g.*, those found on A09, D12), which may be due to biological differences between the sequenced accessions (Fig. S12).

With respect to gene content in each assembly, the number of annotated genes ranged from 70,199 and 87,800 among all of the assemblies studied here, with an average of 75,080. This wide range in annotated gene numbers likely reflects differences in completeness of the genome assemblies and/or differences in annotation pipelines (*e.g.*, software, parameters, *etc.*). To create a cotton-genomics research tool, we conducted genic comparisons between and among the *G. hirsutum* and *G. barbadense* assemblies to identify one-to-one relationships (Tables S13-S15). For the generation of this resource, we selected the genome assemblies TM1_HAU_v1 and 3–79_HAU_v2 as the index (*i.e.*, base reference) as they both were released earlier than other Pacbio+ assemblies and have been used extensively. We tabulated 60,850

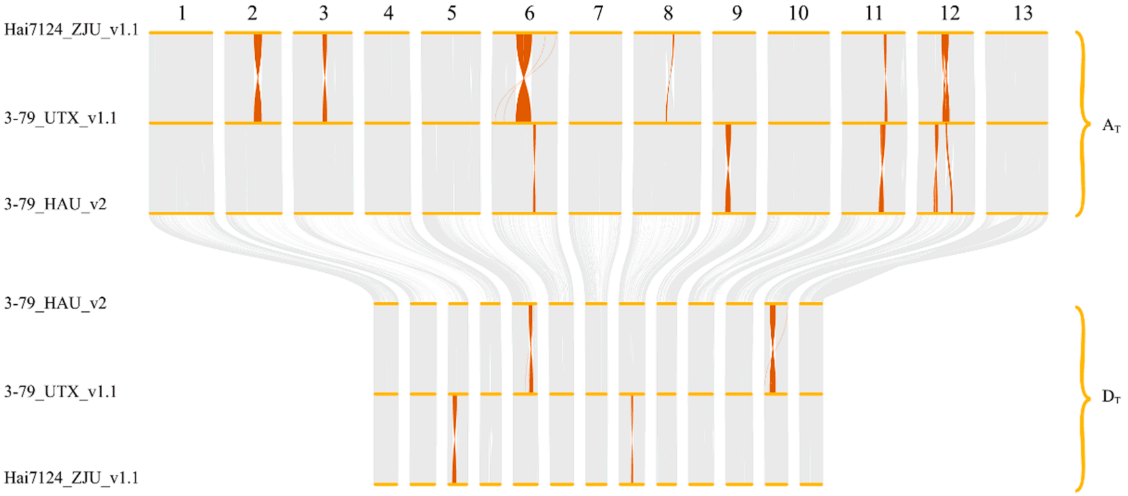

Fig. 6. Genome alignment among three *G. barbadense* genomes, the red lines show the rearrangement sequences larger than 5 Mb.

one-to-one relationships between TM1_HAU_v1 and the other *G. hirsutum* assemblies, and we present these as a side-by-side comparison in Table S13. For *G. barbadense*, there were slightly fewer (59,318) predicted one-to-one relationships between 3 and 79_HAU_v2 and the other *G. barbadense* assemblies (Table S14). Additionally, we also constructed an inter-species, one-to-one relationship map between TM1_HAU_v1 and 3–79_HAU_v2 (Table S15), which may be used in conjunction with the aforementioned *G. hirsutum* and *G. barbadense* gene tables to facilitate gene ID conversions among all of the high-quality domesticated polyploid genomes currently available. To provide a convenient resource for converting gene IDs between different genome versions, we constructed a webpage (https://ihope.shinyapps.io/cottonParalogs/), which allows researchers to leverage previously reported results (based on a separate genome assembly or species) in their research.

3.7. Discussion

During the last twenty years, thousands of genomes for species from non-vascular to flowering plants have been assembled. However, constructing complete genomes and building reference pan-genomes remain important challenges in the plant genomics (Sun et al., 2022). Comparisons among genome assemblies of the same species that were sequenced and assembled using diverse strategies have the potential to identify the best assembly and to help unify future genomics. Our study moves toward this goal with appropriate comparative genomics analyses of multiple assemblies of the two most commercially important cotton genomes. This work should serve as a foundation for the publication of future cotton genome sequences.

This comprehensive assessment found, unsurprisingly, that assemblies sequenced using PacBio long-read sequencing technology were better than those using Illumina short-read sequencing technology and 10X Genomics linked technologies, in terms of sequence contiguity, sequence accuracy, and completeness. Also unsurprisingly, LTR Assembly Index (LAI) scores indicate that assemblies based on PacBio long-reads were better than those based on Illumina short-read sequencing technology alone or even when scaffolded with 10X Genomics technology. The centromeric regions identified in these assemblies using existing ChIP-seq data indicate that the two assemblies sequenced with 10X Genomics identified more chromosomes with centromeric regions than did the assemblies based on PacBio long-reads.

Because scaffolding techniques are essential for building reference-quality assemblies of complex genomes, several technologies have been developed to facilitate accurate scaffolding, including high-throughput chromosome conformation capture (Hi-C) and the BioNano optical mapping. These data can also be used to verify or refute the structural differences that exist among genomes, including the large-scale inversions detected among the five TM-1 assemblies and among the three *G. barbadense* assemblies studied here. Here we used publicly available *G. hirsutum* and *G. barbadense* Hi-C data to evaluate these large-scale inversions for potential assembly artifacts. Although some of the larger inversions (Fig. S9) were shared in the TM1_UTX_v2.1 and TM1_WHU_v1 genomes, these were contraindicated by the Hi-C, suggesting that these inversions are due to the scaffolds being placed in opposite directions in these assemblies rather than biological polymorphism. While most of the inversions among the *G. barbadense* genomes were not supported by the Hi-C data, several inversions between *G. barbadense* 3–79 (Wang et al., 2018) and Hai7124 (Hu et al., 2019) were supported, suggesting that these inversions reflect true differences between the two *G. barbadense* accessions 3–79 and Hai7124 (Fig. S12).

The reference genome quality standard for the Vertebrate Genome Project requires that all released genomes have: (1) an N50 size of at least 1 Mb for contigs and 10 Mb for scaffolds, (2) a sequence error frequency not higher than 1 in 10,000 bases, (3) structural variants that are confirmed by multiple technologies, and (4) at least 90 % of the sequence assigned to chromosomes (Editorial, 2018). In addition, the LTR Assembly Index (LAI) can be used to evaluate assembly continuity using LTR-RTs, noting that genomes with LAI score from 10 to 20 are considered high quality (Ou et al., 2018). Our comprehensive assessment and comparative analysis showed that all tetraploid cotton genome assemblies using long-read sequencing technologies met the reference quality standard. In comparisons among the five high quality TM-1 assemblies, we find that: (1) the BAC alignments indicate that TM1_HAU_v1, TM1_CRI_v1, TM1_UTX_v2.1 and TM1_WHU_v1 have greater coverage and identity than others; (2) BUSCO evaluation indicates general completeness; (3) the contig N50 scores for TM1_WHU_v1, TM1_CRI_v1 and TM1_HAU_v1 are greater than that for TM1_ZJU_v2.1 and TM1_UTX_v2.1 (< 1 Mb); (4) the LAI indicates greater completeness in TM1_WHU_v1 and TM1_CRI_v1; (5) detectable centromeric regions (*via* CenH3 ChIP-seq) are more frequent in TM1_CRI_v1 and TM1_UTX_v2.1; and (6) whole genome alignments indicate multiple artifactual rearrangements in TM1_UTX_v2.1 and TM1_WHU_v1. Accordingly, while the five high-quality TM-1 genomes are generally comparable in metrics, we find that TM1_CRI_v1 is more structurally accurate, indicating that this genome should be preferred when genomic structure is important to the analysis. Similar, for the three *G. barbadense* 3–79 genomes, we find that: (1) BUSCO evaluation indicates slightly more completeness in 3–79_HAU_v2 than 3–79_UTX_v1.1; (2) the contig N50 scores for 3–79_HAU_v2 are greater than that for and 3–79_UTX_v1.1; (3) the LAI indicates greater completeness in 3–79_UTX_v1.1 than 3–79_HAU_v2; (4) detectable centromeric regions (*via* CenH3 ChIP-seq) are more frequent in 3–79_UTX_v1.1 than 3–79_HAU_v2; (5) whole genome alignments indicate multiple artifactual rearrangements in 3–79_UTX_v1.1. Again, this suggests that while any of these genomes are suitable for many analyses, those that consider genome structure may be better served by the 3–79_HAU_v2 assembly.

Accuracy of gene prediction and annotation continues to be a challenge for plant genomics. Many plant genomes have updated their gene prediction and annotation information, such as *Arabidopsis thaliana*, *Oryza sativa*, and *Zea mays* (Cheng et al., 2017; Jiao et al., 2017; Kawahara et al., 2013), using the abundant resources available for these species and some community curation; a similar strategy for updating gene prediction and annotation information would also be useful for the cotton genomes. In the interim, we have generated a resource (https://ihope.shinyapps.io/cottonParalogs/) that allows researchers to convert gene-ids among the different assemblies and leverage information garnered from the myriad assemblies.

Assembling gap-free telomere-to-telomere assemblies is the ideal goal for all-genome sequences. Improvements in the cost and read length of sequencing technologies, and improvements to algorithms of new assemblers, have recently provided telomere-to-telomere complete genomes for the human genome, for two elite *O. sativa* xian/*indica* rice varieties, and for watermelon (Deng et al., 2022; Nurk et al., 2022; Song et al., 2021). The optimization of circular consensus sequencing (CCS) to improve the accuracy of single-molecule real-time (SMRT) sequencing (PacBio) and generate highly accurate long high-fidelity (HiFi) reads, and assembly with the new assemblers, such as HiCanu or Hifiasm, could significantly improve assembly contiguity in cotton (Cheng et al., 2021; Nurk et al., 2020; Perkin et al., 2021; Wenger et al., 2019). While the existing assemblies are high-quality, gap-free telomere-to-telomere assemblies of *G. hirsutum* and *G. barbadense* will likely be achieved in the near future by integrating the latest sequencing technologies (e.g, circular consensus sequencing, BioNano optical mapping, Hi-C, and 10X Genomics). RNA-seq gradually accumulated from different tissues and various classes of non-coding RNA including microRNA, long intergenic RNA, small nucleolar RNA, natural antisense transcript, and small nuclear RNA will also improve the accuracy of the gene predictions and annotation for newly developed genomes. Both the complete genome and accurate gene models will accelerate evolutionary and functional genomic studies in cotton and inform future breeding programs for fiber improvement.

4. Conclusions

This study comprehensively assessed the assembly quality of two cultivated allotetraploid cotton species, *G. hirsutum* and *G. barbadense,* using the same criteria, identified the centromeric regions of all genome assemblies, and revealed one-to-one relationships among genes within and between species, the latter providing a convenient resource for transferring gene IDs between species and among genome versions. Several large inversions among the five high-quality TM-1 assemblies reveals the nonconformity among independently generated genome sequences using similar technologies. Our analyses collectively reveal that the assembly quality of TM1_CRI_v1 is slightly higher than other *G. hirsutum* assemblies, and the assembly quality of 3–79_HAU_v2 is slightly higher than other *G. barbadense* assemblies. Our analysis provides a path forward for choosing the most appropriate reference genome for cotton research and provides resources for cotton improvement.

Funding

This work was financially supported by National Key Research and Development Plan (NO. 2021YFF1000102) and the Fundamental Research Funds for the Central Universities (NO. 2021ZKPY013) to Daojun Yuan.

CRediT authorship contribution statement

Qingying Meng: Data curation, Formal analysis, Visualization, Validation, Writing – original draft, Writing – review & editing. **Jiaqi Gu:** Software, Writing – review & editing. **Zhongping Xu:** Software. **Jie Zhang:** Writing – review & editing. **Jiwei Tang:** Writing – review & editing. **Anzhou Wang:** Formal analysis. **Ping Wang:** Investigation. **Zhaowei Liu:** Validation. **Yuxuan Rong:** Data curation. **Peihao Xie:** Investigation. **Liuyang Hui:** Data curation. **Joshua A. Udall:** Writing – review & editing. **Corrinne E. Grover:** Writing – review & editing. **Jonathan F. Wendel:** Writing – review & editing. **Shuangxia Jin:** Software. **Xianlong Zhang:** Conceptualization, Methodology. **Daojun Yuan:** Conceptualization, Methodology, Validation, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgement

The computations in this paper were run on the bioinformatics computing platform of the National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.indcrop.2023.116471.

References

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Cai, Y., Cai, X., Wang, Q., 2020. Genome sequencing of the Australian wild diploid species *Gossypium australe* highlights disease resistance and delayed gland morphogenesis. Plant Biotechnol. J. 18, 814–828. https://doi.org/10.1111/pbi.13249.

Chen, Z.J., Scheffler, B.E., Dennis, E., 2007. Toward sequencing cotton (*Gossypium)* Genomes. Plant Physiol. 145, 1303. https://doi.org/10.1104/pp.107.107672.

Chen, Z.J., Sreedasyam, A., Ando, A., 2020. Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. Nat. Genet. 52, 525–533. https://doi.org/10.1038/s41588-020-0614-5.

Cheng, C.-Y., Krishnakumar, V., Chan, A.P., 2017. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. Plant J. 89, 789–804. https://doi.org/10.1111/tpj.13415.

Cheng, H., Concepcion, G.T., Feng, X., 2021. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. Nat. Methods 18, 170–175. https://doi.org/10.1038/s41592-020-01056-5.

Delcher, A.L., Salzberg, S.L., Phillippy, A.M., 2003. Using MUMmer to identify similar regions in large sequence sets. Curr. Protoc. Bioinform. https://doi.org/10.1002/0471250953.bi1003s00.

Deng, Y., Liu, S., Zhang, Y., 2022. A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene discovery and breeding. Mol. Plant 15, 1268–1284. https://doi.org/10.1016/j.molp.2022.06.010.

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

Durand, N.C., Robinson, J.T., Shamim, M.S., 2016. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 3, 99–101. https://doi.org/10.1016/j.cels.2015.07.012.

Editorial, 2018. A reference standard for genome biology. Nat. Biotechnol. 36, 1121. https://doi.org/10.1038/nbt.4318.

Ellinghaus, D., Kurtz, S., Willhoeft, U., 2008. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. BMC Bioinform. 9, 18. https://doi.org/10.1186/1471-2105-9-18.

Endrizzi, J.E., Turcotte, E.L., Kohel, R.J., 1985. Genetics, cytology, and evolution of *Gossypium*. In: Caspari, E.W., Scandalios, J.G. (Eds.), Advances in Genetics. Academic Press, pp. 271–375.

Gallagher, J.P., Grover, C.E., Rex, K., 2017. A new species of cotton from Wake Atoll, *Gossypium Stephensii* (*Malvaceae*). Syst. Bot. 42 (1), 115–123. https://doi.org/10.1600/036364417×694593.

Goel, M., Sun, H., Jiao, W.-B., 2019. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. 20, 277. https://doi.org/10.1186/s13059-019-1911-0.

Grover, C.E., Gallagher, J.P., Jareczek, J.J., 2015. Re-evaluating the phylogeny of allopolyploid *Gossypium* L. Mol. Phylogenet. Evol. 92, 45–52. https://doi.org/10.1016/j.ympev.2015.05.023.

Grover, C.E., Pan, M., Yuan, D., 2020. The *Gossypium longicalyx* genome as a resource for cotton breeding and evolution. G3 Genes Genomes Genet. 10, 1457–1467. https://doi.org/10.1534/g3.120.401050.

Grover, C.E., Yuan, D., Arick II, M.A., 2021a. The *Gossypium anomalum* genome as a resource for cotton improvement and evolutionary analysis of hybrid incompatibility. G3 Genes Genomes Genet. 11, jkab319 https://doi.org/10.1093/g3journal/jkab319.

Grover, C.E., Yuan, D., Arick II, M.A., 2021b. The *Gossypium stocksii* genome as a novel resource for cotton improvement. G3 Genes Genomes Genet. 11, jkab125 https://doi.org/10.1093/g3journal/jkab125.

Guo, W., Cai, C., Wang, C., 2008. A preliminary analysis of genome structure and composition in *Gossypium hirsutum*. BMC Genom. 9, 314. https://doi.org/10.1186/1471-2164-9-314.

Han, J., Masonbrink, R.E., Shan, W., 2016. Rapid proliferation and nucleolar organizer targeting centromeric retrotransposons in cotton. Plant J. 88, 992–1005. https://doi.org/10.1111/tpj.13309.

He, S., Sun, G., Geng, X., 2021. The genomic basis of geographic differentiation and fiber improvement in cultivated cotton. Nat. Genet. 53, 916–924. https://doi.org/10.1038/s41588-021-00844-9.

Hendrix, B., Stewart, J.M., 2005. Estimation of the nuclear DNA content of *Gossypium* species. Ann. Bot. 95, 789–797. https://doi.org/10.1093/aob/mci078.

Hu, Y., Chen, J., Fang, L., 2019. *Gossypium barbadense* and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton. Nat. Genet. 51, 739–748. https://doi.org/10.1038/s41588-019-0371-5.

Huang, G., Wu, Z., Percy, R.G., 2020. Genome sequence of *Gossypium herbaceum* and genome updates of *Gossypium arboreum* and *Gossypium hirsutum* provide insights into cotton A-genome evolution. Nat. Genet. https://doi.org/10.1038/s41588-020-0607-4.

Jiao, Y., Peluso, P., Shi, J., 2017. Improved maize reference genome with single-molecule technologies. Nature 546, 524–527. https://doi.org/10.1038/nature22971.

Kawahara, Y., de la Bastide, M., Hamilton, J.P., 2013. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice 6, 4. https://doi.org/10.1186/1939-8433-6-4.

Kent, W.J., 2002. BLAT–the BLAST-like alignment tool. Genome Res. 12, 656–664. https://doi.org/10.1101/gr.229202.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923.

Li, F., Fan, G., Lu, C., 2015. Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. Nat. Biotechnol. 33, 524–530. https://doi.org/10.1038/nbt.3208.

Li, F., Fan, G., Wang, K., 2014. Genome sequence of the cultivated cotton *Gossypium arboreum*. Nat. Genet. 46, 567–572. https://doi.org/10.1038/ng.2987.

Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Li, H., Handsaker, B., Wysoker, A., 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.

Liu, B., Brubaker, C.L., Mergeai, G., 2001. Polyploid formation in cotton is not accompanied by rapid genomic changes. Genome 44, 321–330. https://doi.org/10.1139/g01-011.

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

Ma, Z., Zhang, Y., Wu, L., 2021. High-quality genome assembly and resequencing of modern cotton cultivars provide resources for crop improvement. Nat. Genet. 53, 1385–1391. https://doi.org/10.1038/s41588-021-00910-2.

Nurk, S., Koren, S., Rhie, A., 2022. The complete sequence of a human genome. Science 376, 44–53. https://doi.org/10.1126/science.abj6987.

Nurk, S., Walenz, B.P., Rhie, A., 2020. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res. 30, 1291–1305. https://doi.org/10.1101/gr.263566.120.

Ou, S., Chen, J., Jiang, N., 2018. Assessing genome assembly quality using the LTR Assembly Index (LAI. Nucleic Acids Res. 46, e126-e126 https://doi.org/10.1093/nar/gky730.

Ou, S., Jiang, N., 2018. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat Retrotransposons. Plant Physiol. 176, 1410. https://doi.org/10.1104/pp.17.01310.

Ou, S., Jiang, N., 2019. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. Mob. DNA 10, 48. https://doi.org/10.1186/s13100-019-0193-0.

Parkinson, J., Blaxter, M., 2009. Expressed sequence tags: an overview. In: Parkinson, J. (Ed.), Expressed Sequence Tags (ESTs): Generation and Analysis. Humana Press, Totowa, NJ, pp. 1–12.

Paterson, A.H., Wendel, J.F., Gundlach, H., 2012. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature 492, 423–427. https://doi.org/10.1038/nature11798.

Peng, R., Xu, Y., Tian, S., 2022. Evolutionary divergence of duplicated genomes in newly described allotetraploid cottons. Proc. Natl. Acad. Sci. USA, 119, e2208496119. https://doi.org/10.1073/pnas.2208496119.

Perkin, L.C., Bell, A., Hinze, L.L., 2021. Genome assembly of two nematode-resistant cotton lines (Gossypium hirsutum L.). G3 Genes Genomes Genet. 11, jkab276 https://doi.org/10.1093/g3journal/jkab276.

Ramaraj, T., Grover, C.E., Mendoza, A.C., 2022. The *Gossypium herbaceum* L. Wagad genome as a resource for understanding cotton domestication (2022.2006.2007.494775). bioRxiv https://doi.org/10.1101/2022.06.07.494775.

Saski, C.A., Scheffler, B.E., Hulse-Kemp, A.M., 2017. Sub genome anchored physical frameworks of the allotetraploid Upland cotton (*Gossypium hirsutum* L.) genome, and an approach toward reference-grade assemblies of polyploids. Sci. Rep. 7, 15274. https://doi.org/10.1038/s41598-017-14885-w.

Schäffer, A.A., Aravind, L., Madden, T.L., 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. 29, 2994–3005. https://doi.org/10.1093/nar/29.14.2994.

Schäffer, A.A., Nawrocki, E.P., Choi, Y., 2017. VecScreen_plus_taxonomy: imposing a tax (onomy) increase on vector contamination screening. Bioinformatics 34, 755–759. https://doi.org/10.1093/bioinformatics/btx669.

Servant, N., Varoquaux, N., Lajoie, B.R., 2015. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 16, 259. https://doi.org/10.1186/s13059-015-0831-x.

Sheng, K., Sun, Y., Liu, M., 2022. A reference-grade genome assembly for *Gossypium bickii* and insights into its genome evolution and formation of pigment gland and gossypol. Plant Commun., 100421 https://doi.org/10.1016/j.xplc.2022.100421.

Song, J.-M., Xie, W.-Z., Wang, S., 2021. Two gap-free reference genomes and a global view of the centromere architecture in rice. Mol. Plant. https://doi.org/10.1016/j.molp.2021.06.018.

Sun, Y., Shang, L., Zhu, Q.-H., 2022. Twenty years of plant genome sequencing: achievements and challenges. Trends Plant Sci. 27, 391–401. https://doi.org/10.1016/j.tplants.2021.10.006.

Udall, J.A., Long, E., Hanson, C., 2019. De Novo genome sequence assemblies of *Gossypium raimondii* and *Gossypium turneri*. G3 9, 3079–3085. https://doi.org/10.1534/g3.119.400392.

Wang, K., Wang, Z., Li, F., 2012a. The draft genome of a diploid cotton *Gossypium raimondii*. Nat. Genet. 44, 1098–1103. https://doi.org/10.1038/ng.2371.

Wang, L., Wang, G., Long, L., 2020. Understanding the role of phytohormones in cotton fiber development through omic approaches; recent advances and future directions. Int. J. Biol. Macromol. 163, 1301–1313. https://doi.org/10.1016/j.ijbiomac.2020.07.104.

Wang, M., Li, J., Wang, P., 2021. Comparative genome analyses highlight transposon-mediated genome expansion and the evolutionary architecture of 3D genomic folding in cotton. Mol. Biol. Evol. https://doi.org/10.1093/molbev/msab128.

Wang, M., Tu, L., Yuan, D., 2019. Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. Nat. Genet. 51, 224–229. https://doi.org/10.1038/s41588-018-0282-x.

Wang, M., Wang, P., Lin, M., 2018. Evolutionary dynamics of 3D genome architecture following polyploidization in cotton. Nat. Plants 4, 90–97. https://doi.org/10.1038/s41477-017-0096-3.

Wang, Y., Tang, H., DeBarry, J.D., 2012b. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40, e49-e49 https://doi.org/10.1093/nar/gkr1293.

Wendel, J., Grover, C., 2015. Taxonomy and Evolution of the Cotton Genus, *Gossypium*.

Wendel, J.F., 1989. New World tetraploid cottons contain Old World cytoplasm. Proc. Natl. Acad. Sci. USA, 86, 4132–4136. https://doi.org/10.1073/pnas.86.11.4132.

Wenger, A.M., Peluso, P., Rowell, W.J., 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol. 37, 1155–1162. https://doi.org/10.1038/s41587-019-0217-9.

Xu, H., Luo, X., Qian, J., 2012. FastUniq: a fast *de novo* duplicates removal tool for paired short reads. PLoS One 7, e52249. https://doi.org/10.1371/journal.pone.0052249.

Xu, Z., Chen, J., Meng, S., 2022. Genome sequence of *Gossypium anomalum* facilitates interspecific introgression breeding. Plant Commun., 100350 https://doi.org/10.1016/j.xplc.2022.100350.

Yang, N., Liu, J., Gao, Q., 2019a. Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nat. Genet. 51, 1052–1059. https://doi.org/10.1038/s41588-019-0427-6.

Yang, Z., Ge, X., Yang, Z., 2019b. Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. Nat. Commun. 10, 2989. https://doi.org/10.1038/s41467-019-10820-x.

Yu, J., Jung, S., Cheng, C.-H., 2021. CottonGen: the community database for cotton genomics, genetics, and breeding research. Plants 10. https://doi.org/10.3390/plants10122805.

Yuan, D., Tang, Z., Wang, M., 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. https://doi.org/10.1038/srep17662.

Zang, C., Schones, D.E., Zeng, C., 2009. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics 25, 1952–1958. https://doi.org/10.1093/bioinformatics/btp340.

Zhang, T., Hu, Y., Jiang, W., 2015. Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. Nat. Biotechnol. 33, 531–537. https://doi.org/10.1038/nbt.3207.