The Gossypium longicalyx Genome as a Resource for Cotton Breeding and Evolution

Corrinne E. Grover,* Mengqiao Pan,† Daojun Yuan,‡ Mark A. Arick,§ Guanjing Hu,* Logan Brase,** David M. Stelly,†† Zefu Lu,§ Robert J. Schmitz,‡ Daniel G. Peterson,§ Jonathan F. Wendel,* and Joshua A. Udall†††,8,†

*Ecology, Evolution, and Organismal Biology Dept., Iowa State University, Ames, IA, 50010, †State Key Laboratory of Crop Genetics and Germplasm Enhancement, Cotton Hybrid R & D Engineering Center, Nanjing Agricultural University, Nanjing, 210095, China, ‡College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, Hubei, 430070, China, §Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, **Division of Biology and Biomedical Sciences, Washington University in St. Louis, St. Louis, MO 63110, ††Department of Soil and Crop Sciences, Texas A&M University, College Station, 77843, †‡USDA/Agricultural Research Service, Crop Germplasm Research Unit, College Station, TX 77845, and §§Department of Genetics, University of Georgia, Athens, GA 30602

ORCID IDs: 0000-0003-3878-5459 (C.E.G.); 0000-0001-6007-5571 (D.Y.); 0000-0002-7207-3052 (M.A.A.); 0000-0001-8552-7394 (G.H.); 0000-0002-7175-3208 (L.B.); 0000-0002-3468-4119 (D.M.S.); 0000-0001-7538-6663 (R.J.S.); 0000-0002-0274-5968 (D.G.P.); 0000-0003-2258-5081 (J.F.W.); 0000-0003-0978-4764 (J.A.U.)

ABSTRACT Cotton is an important crop that has made significant gains in production over the last century. Emerging pests such as the reniform nematode have threatened cotton production. The rare African diploid species Gossypium longicalyx is a wild species that has been used as an important source of reniform nematode immunity. While mapping and breeding efforts have made some strides in transferring this immunity to the cultivated polyploid species, the complexities of interploidal transfer combined with substantial linkage drag have inhibited progress in this area. Moreover, this species shares its most recent common ancestor with the cultivated A-genome diploids, thereby providing insight into the evolution of long, spinnable fiber. Here we report a newly generated de novo genome assembly of G. longicalyx. This high-quality genome leveraged a combination of PacBio long-read technology, Hi-C chromatin conformation capture, and BioNano optical mapping to achieve a chromosome level assembly. The utility of the G. longicalyx genome for understanding reniform immunity and fiber evolution is discussed.

KEYWORDS Gossypium longicalyx nematode resistance cotton fiber genome sequence PacBio

Cotton (genus Gossypium) is an important crop that provides the largest natural source of fiber. Colloquially, the term cotton refers to one of four domesticated species, primarily the tetraploid G. hirsutum (n = 26), which is responsible for over 98% of cotton production worldwide (Kranthi 2018). Gossypium contains over 50 additional wild species related to the domesticated cottons (partitioned into groups of closely related species designated “A-G” and “K” for diploids; “AD” for tetraploids), which serve as potential sources of disease and pest resistance. Among these, Gossypium longicalyx J.B. Hutch. & B.J.S. Lee (n = 13) is the only representative of the diploid F-genome (Wendel and Grover 2015) and the only species with immunity to reniform nematode infection (Yik and Birchfield 1984). Discovered only 60 years ago (Hutchinson and B. J. S. Lee 1958), it is both cytogenetically differentiated from members of the other genome groups (Phillips 1966) and morphologically isolated (Fryxell 1971, 1992). Importantly, G. longicalyx is sister to the A-genome cottons (Wendel and Albert 1992; Wendel and Grover 2015; Chen et al. 2016), i.e., G. arboreum and G. herbaceum (both n = 13), the only diploids with long, spinnable fiber.

Interest in the genome of G. longicalyx is twofold. First, broad-scale screening of the cotton germplasm collection indicates that domesticated cotton lacks appreciable natural resistance to reniform
nematode (Birchfield and Brister 1963; Yik and Birchfield 1984), and while several other species exhibit degrees of resistance, only *G. longicalyx* exhibits immunity to infection (Yik and Birchfield 1984). This is significant as reniform nematode has emerged as a major source of cotton crop damage, reducing cotton production by over 205 million bales per year (Lawrence et al. 2015) and accounting for ~11% of the loss attributable to pests (Khanal et al. 2018). Current reniform resistant lines are derived from complex breeding schemes which are required to introgress reniform immunity from the diploid *G. longicalyx* into polypl oid *G. hirsutum* (Bell and Robinson 2004; Digh et al. 2009; Khanal et al. 2018); however, undesirable traits have accompanied this introgression (Nichols et al. 2010) extreme stunting of seedlings and plants exposed to dense nematode populations, prohibiting commercial deployment (Zheng et al. 2016).

The genome of *G. longicalyx* is also valuable because it is phylogenetically sister to the only diploid clade with spinnable fiber (Wendel and Albert 1992; Wendel and Grover 2015; Chen et al., 2016), the A-genome species, which contributed the maternal ancestor to polypl oid cotton. Consequently, there has been interest in this species as the ancestor to spinnable fiber (Hovav et al. 2008; Paterson et al. 2012), although progress has been limited due to lack of genomic resources in *G. longicalyx*. Comparisons between the *G. longicalyx* genome and other cotton genomes, including the domesticated diploids (Du et al. 2018), may provide clues into the evolutionary origin of “long” fiber.

Here we describe a high-quality, de novo genome sequence for *G. longicalyx*, a valuable resource for understanding nematode immunity in cotton and possibly other species. This genome also provides a foundation to understand the evolutionary origin of spinnable fiber in *Gossypium*.

**METHODS & MATERIALS**

**Plant material and sequencing methods**

Leaf tissue of mature *G. longicalyx* (F1-1) was collected from a Brigham Young University (BYU) greenhouse. DNA was extracted using CTAB techniques (Kidwell and Osborn 1992), and the amount recovered was measured via Qubit Fluorometer (ThermoFisher, Inc.). The sequencing library was constructed by the BYU DNA Sequencing Center (DNASC) using only fragments >18 kb, which were size selected on the BluePippen (Sage Science, LLC) and verified in size using a Fragment Analyzer (Advanced Analytical Technologies, Inc.). Twenty-six PacBio cells were sequenced from a single library on the Pacific Biosciences Sequel system. Resulting reads were assembled using Canu V.1.6 using default parameters (Koren et al. 2017) to create a sequence assembly called Longicalyx_V1.0, composed of 229 large contigs (Supplemental Figure 1). This draft assembly was combined with Chicago HiRise (Koch et al. 2016; Putnam et al. 2016) + Dovetail Hi-C libraries by Dovetail Genomics to produce an intermediate assembly, Longicalyx_V3.0.

High-molecular weight DNA was extracted from young *G. longicalyx* leaves and subsequently purified, nicked, labeled, and repaired according to Bionano Plant protocol and standard operating procedures for the Irys platform. BssSI was used in conjunction with the IrysSolve pipeline to assemble an optical map on the BYU Fulton SuperComputing cluster. The resulting optical map was aligned to the assembly named Longicalyx_V3.0 using an in silico labeled reference sequence. Bionano maps linked large contigs present in this assembly, producing 17 large scaffolds (Longicalyx_V4.0).

Minion sequencing libraries were created and sequenced following the standard protocol from Oxford Nanopore. Scaffolds from Longicalyx_V4.0 were polished (Supplemental File 1) with existing Illumina (SRR1174179 and SRR1174182 from the NCBI Short Read Archive) and the newly generated Minion data for *G. longicalyx* using both PBjelly (English et al. 2012) and GapFiller (Boetzer and Pirovano 2012) to produce the final assembly, Longicalyx_V5.0.

**Repeat and gene annotation**

Repeats were identified using two methods. The first is a homology-based approach, i.e., a combination of RepeatMasker (Smit et al. 2015) and “One code to find them all” (Bailey-Beth et al. 2014), whereas the second method (i.e., RepeatExplorer; Novak et al. 2010) clusters reads based on sequence similarity and automatically annotates the most abundant cluster using RepeatMasker. Each RepeatMasker run used a custom library, which combines Repbase 23.04 repeats (Bao et al. 2015) with cotton-specific repeats. Default parameters were run, except the run was “sensitive” and was set to mask only TE (no low-complexity). Parameters are available https://github.com/Wendelab/longicalyx. “One code to find them all” was used to aggregate multiple hits from the first method (RepeatMasker) into TE models using default parameters. The resulting output was aggregated and summarized in R/3.6.0 (R Core Team 2017) using dplyr 0.8.1(Wickham et al. 2015). Cluster results were obtained from (Grover et al. 2019) and https://github.com/IGIBD_D_Cottons_USDA, and these were parsed in R/3.6.0 (R Core Team 2017). All code is available at https://github.com/Wendelab/longicalyx.

RNA-Seq libraries were generated from *G. longicalyx* leaf (CL), floral (FF), and stem tissues (FS) to improve genome annotation. RNA-seq libraries were independently constructed by BGI Americas (Davis, CA) using Illumina TruSeq reagents and subsequently sequenced (single-end, 50 bp). The newly sequenced *G. longicalyx* RNA-seq was combined with existing RNA-seq from *G. longicalyx* (SRR1174179) as well as two closely related species, i.e., *G. herbaceum* (developing fibers and seed; PRJNA595350 and SRR959585, respectively) and *G. arboreum* (5 seed libraries and 1 seedling; SRR617075, SRR617073, SRR617068, SRR617067, SRR959590, and SRR959508). RNA-seq libraries were mapped to the hard-masked *G. longicalyx* genome using hisat2 v2.1.0 (Kim et al. 2015). BRAKER2 v2.1.2 (Hoff et al. 2019) was used in conjunction with GeneMark [v4.36 (Borodovsky and Lomsadze 2011) was used to produce high quality RNA-seq based gene predictions by combining the RNA-seq assemblies produced by StringTie [v1.3.6] (Pertea et al. 2015) and Cufflinks [v2.2.1] (Ghosh and Chan 2016) with a reference-guided assembly from Trinity [v2.8.5] (Grabherr et al. 2011) and a splice junction analysis from Portcullis [v1.2.2] (Mapleson et al. 2018). The Trinity assembly was formatted using GMAP [v2019-05-12] (Wu and Watanabe 2005). MAKER2 [v2.31.10] (Holt and Yandell 2011; Campbell et al. 2014) was used to integrate gene predictions from (1) BRAKER2 trained Augustus, (2) GeneMark, and (3) Mikado, also using evidence from all Gossypium ESTs available from NCBI (nucleotide database filtered on “txid3633” and “is_est”) and a database composed of all curated proteins in Uniprot Swissprot [v2019_07] (UniProt Consortium 2008) combined with the annotated proteins from the *G. hirsutum* (https://www.cottongen.org/species/Gossypium_hirsutum/igi-AD1_genome_v1.1) and *G. raimondii* (n = 13; Paterson et al. 2012) genomes. Maker scored each gene model using the annotation edit distance (AED) - (Eilbeck et al. 2009; Holt and Yandell 2011; Yandell and Ence 2012) metric based on EST and protein evidence provided. Gene models with an AED greater than 0.47 were removed from further analyses, and the

---

C. E. Grover et al.
remaining gene models were functionally annotated using InterProScan [v5.35-74.0] (Jones et al. 2014) and BlastP [v2.9.0+] (Camacho et al. 2009) searches against the Uniprot SwissProt database. Orthologs between the G. longicalyx annotations and the existing annotations for G. arboreum (Du et al. 2018), G. raimondii (Paterson et al. 2012), G. hirsutum (Hu et al. 2019), and G. barbadense (n = 26; Hu et al. 2019) were predicted by OrthoFinder using default settings (Emms and Kelly 2015, 2019). All genomes are hosted through CottonGen (https://www.cottongen.org Yu et al. 2014) and running parameters are available from https://github.com/Wendellab/longicalyx.

**ATAC-seq and data analysis**

ATAC-seq was performed as described previously (Lu et al. 2017). For each replicate, approximately 200 mg freshly collected leaves or flash frozen leaves were immediately chopped with a razor blade in ~1 ml of pre-chilled lysis buffer (15 mM Tris-HCl pH 7.5, 20 mM NaCl, 80 mM KCl, 0.5 mM spermine, 5 mM 2-mercaptoethanol, 0.2% Triton X-100). The chopped slurry was filtered twice through miracloth and once through a 40 µm filter. The crude nuclei were stained with DAPI and loaded into a flow cytometer (Beckman Coulter MoFlo XDP). Nuclei were purified by flow sorting and washed in accordance with Lu et al. (Lu et al. 2017). The sorted nuclei were incubated with 2 µl Tn5 transposomes in 40 µl of tagmentation buffer (10 mM TAPS-NaOH ph 8.0, 5 mM MgCl2) at 37° for 30 min without rotation. The integration products were purified using a Qiagen MinElute PCR Purification Kit or NEB Monarch DNA Cleanup Kit and then amplified using Phusion DNA polymerase for 10-13 cycles. PCR cycles were determined as described previously (Buenrostro et al. 2013). Amplified libraries were purified with AMPure beads to remove primers. ATAC-seq libraries were sequenced in paired-end 35 bp at the University of Georgia Genomics & Bioinformatics Core using an Illumina NextSeq 500 instrument.

Reads were adapter and quality trimmed, and then filtered using “Trim Galore” [v0.4.5] (Krueger 2015). Clean reads were subsequently aligned to the Longicalyx_V5.0 assembly using Bowtie2 [v2.3.4] (Langmead and Salzberg 2012) with the parameters “–no-mixed–no-discardant–no-unal–dovetail”. Duplicate reads were removed using Picard [v2.17.0] with default parameters (http://broadinstitute.github.io/picard/). Only uniquely mapped read pairs with a quality score of at least 20 were kept for peak calling. Phantompeakaktoolits [v1.14] (Landt et al. 2012) was used to calculate the strand cross-correlation, and deepTools [v2.5.2] (Ramirez et al. 2016) was used to calculate correlation between replicates. The peak calling tool from HOMER [v4.10] (Heinz et al. 2010), i.e., findpeaks, was run in “region” mode and with the minimum distance between peaks set to 150 bp. MACS2 [v2.1.1] (Zhang et al. 2008) callpeak, a second peak-calling algorithm, was run with the parameter “-f BAMPE” to analyze only properly paired alignments, and putative peaks were filtered using default settings and false discovery rate (FDR) < 0.05. Due to the high level of mapping reproducibility (Pearson’s correlation r = 0.99 and Spearman correlation r = 0.77 by deepTools), peaks were combined and merged between replicates for each tool using BEDTools [v2.27.1] (Quinlan 2014). BEDTools was also used to intersect HOMER peaks and MACS2 peaks to only retain peak regions identified by both tools as accessible chromatin regions (ACRs) for subsequent analyses.

ACRs were annotated in relation to the nearest annotated genes in the R environment [v3.5.0] as genic (gACRs; overlapping a gene), proximal (pACRs; within 2 Kb of a gene) or distal (dACRs; > 2 Kb from a gene). Using R package ChIPseeker [v1.18.0] (Yu et al. 2015), the distribution of ACRs was calculated around transcription start sites (TSS) and transcription termination sites (TTS), and peak distribution was visualized with aggregated profiles and heatmaps. To compare GC contents between ACRs and non-accessible genomic regions, the BEDTools shuffle command was used to generate the distal (by excluding genic and 2 Kb flanking regions) and genic/proximal control regions (by including genic and 2 Kb flanking regions), and the nuc command was used to calculate GC content for each ACR and permuted control regions.

**Identification of the RenLon region in G. longicalyx**

Previous research (Dighe et al. 2009; Zheng et al. 2016) identified a marker (BNL1231) that consistently cosegregates with resistance and that is flanked by the SNP markers Gl_168758 and Gl_072641, which are all located in the region of G. longicalyx chromosome 11 referred to as “RenLon”. These three markers were used as queries of gmap (Wu and Watanabe 2005) against the assembled genome to identify the genomic regions associated with each. The coordinates identified by gmap were placed in a bed file; this file was used in conjunction with the G. longicalyx annotation and bedtools intersect (Quinlan 2014) to identify predicted G. longicalyx genes contained within RenLon. Samtools faiidx (Li et al. 2009) was used to extract the 52 identified genes from the annotation file, which were functionally annotated using blast2go (blast2go basics; biobam) and including blastx (Altschul et al. 1990), gene ontology (The Gene Ontology Consortium 2019), and InterPro (Jones et al. 2014). Orthogroups containing each of the 52 RenLon genes were identified from the Orthofinder results (see above).

**Comparison between G. arboreum and G. longicalyx for fiber evolution**

Whole-genome alignments were generated between G. longicalyx and either G. arboreum, G. raimondii, G. turneri (Udall et al. 2019), G. hirsutum (A-chromosomes), and G. barbadense (A-chromosomes) using Mummer (MaCris et al. 2018) and visualized using dotPlotly (https://github.com/tpoorten/dotPlotly) in R (version 3.6.0) (R Development Core Team and Others 2011). Divergence between G. longicalyx and G. arboreum or G. raimondii was calculated using orthogroups that contain a single G. longicalyx gene with a single G. arboreum and/or single G. raimondii gene. Pairwise alignments between G. longicalyx and G. arboreum or G. raimondii were generated using the linsi from MAFFT (Katoh and Standley 2013). Pairwise distances between G. longicalyx and G. arboreum and/or G. raimondii were calculated in R (version 3.6.0) using phangorn (Schliep 2011) and visualized using ggsplot2 (Wickham 2016). To identify genes unique to species with spinnable fiber (i.e., G. arboreum and the polyploid species), we extracted any G. arboreum gene contained within orthogroups composed solely of G. arboreum or polyploid A-genome gene annotations, and subjected these to blast2go (as above). Syntenic conservation of genes contained within the RenLon region, as compared to G. arboreum, was evaluated using GEvo as implemented in SynMap via COGE (Lyons and Freeling 2008; Haug-Baltzell et al. 2017).

**Data availability**

The assembled genome sequence of G. longicalyx is available at NCBI under PRJNA420071 and CottonGen (https://www.cottongen.org/). The raw data for G. longicalyx are also available at NCBI PRJNA420071 for PacBio and Minion, and PRJNA420070 for RNA-Seq. Supplemental material available at figshare: https://doi.org/10.25387/g3.11865921.
### RESULTS AND DISCUSSION

#### Genome assembly and annotation

We report a *de novo* genome sequence for *G. longicalyx*. This genome was first assembled from ~144x coverage (raw) of PacBio reads, which alone produced an assembly consisting of 229 contigs with an N50 of 28.8 Mb (Table 1). The contigs were scaffolded using a combination of Chicago Highrise, Hi-C, and BioNano to produce a chromosome level assembly consisting of 17 contigs with an average length of 70.4 Mb (containing only 8.4 kb of gap sequence). Thirteen of the chromosomes were assembled into single contigs. Exact placement of the three unscaffolded contigs (~100 kb) was not determined, but these remaining sequences were included in NCBI with the assembled chromosomes. The final genome assembly size was 1190.7 Mb, representing over 90% of the estimated genome size (Hendrix and Stewart 2005).

To assess genome assembly, we performed a BUSCO analysis of the completed genome (Waterhouse et al. 2017), which recovered 95.8% complete BUSCOs (from the total of 2121 BUSCO groups searched; Table 2). Most BUSCOs (86.5%) were both complete and single copy, with only 9.3% BUSCOs complete and duplicated. Less than 5% of BUSCOs were either fragmented (1.4%) or missing (2.8%), indicating a general completeness of the genome. Genome contiguity was independently verified using the LTR Assembly Index (LAI) (On et al. 2018), which is a reference-free method to assess genome contiguity by evaluating the completeness of LTR-retrotransposon assembly within the genome. This method, applied to over 100 genomes in Phytozome, suggests that an LAI between 10 and 20 should be considered “reference-quality”; the *G. longicalyx* genome reported here received an LAI score of 10.74. Comparison of the *G. longicalyx* genome to published cotton genomes (Table 2) suggests that the quality of this assembly is similar or superior to other currently available cotton genomes.

Genome annotation produced 40,181 transcripts representing 38,378 unique genes. Comparatively, the reference sequences for the related diploids *G. raimondii* (Paterson et al. 2012) and *G. arboreum* (Du et al. 2018) recovered 37,505 and 40,960 genes, respectively. Ortholog analysis between *G. longicalyx* and both diploids suggests a simple 1:1 relationship between a single *G. longicalyx* gene and a single *G. raimondii* or *G. arboreum* gene for 67–68% of the *G. longicalyx* genes (25,637 and 26,249 genes, respectively, out of 38,378 genes; Table 3). Approximately 7–8% of the *G. longicalyx* genome (*i.e.*, 2,615 and 3,158 genes) are in “one/many” (Table 3) relationships whereby one or more *G. longicalyx* gene model(s) matches one or more *G. raimondii* or *G. arboreum* gene model(s), respectively. The remaining 5,009 genes were not placed in orthogroups with any other cotton genome, slightly higher than the 2,016–2,556 unplaced genes in the other diploid species used here. While this could be partly due to genome annotation differences in annotation pipelines, it is also likely due to differences in the amount of RNA-seq available for each genome.

#### Repeats

Transposable element (TE) content was predicted for the genome, both by *de novo* TE prediction (Bailly-Bechet et al. 2014; Smit et al. 2015) and repeat clustering (Novák et al. 2010). Between 44–50% of the *G. longicalyx* genome is inferred to be repetitive by RepeatMasker and RepeatExplorer, respectively. While estimates for TE categories (*e.g.*, DNA, Ty3/gypsy, Ty1/copia, etc.) were reasonably consistent between the two methods (Supplemental Table 1), RepeatExplorer recovered nearly 100 additional megabases of putative repetitive sequences, mostly in the categories of Ty3/gypsy, unspecified LTR elements, and unknown repetitive elements. Interestingly, Repeat-Masker recovered a greater amount of sequence attributable to Ty1/copia and DNA elements (Supplemental Table 1); however, this only accounted for a total of 22 Mb (less than 20% of the total differences over all categories). The difference between methods with respect to each category and the total TE annotation is relatively small and may be attributable to a combination of methods (homology-based TE identification method vs. similarity clustering), the under-exploration

---

**Table 1** Statistics for assembly versions

| Method                  | PacBio/Canu | +Chicago HighRise+HiC | +BioNano | +Illumina+Minion |
|-------------------------|-------------|-----------------------|----------|------------------|
| Coverage                | 79.45       | 79.59                 | 80.47    | 80.42            |
| Total Contig Number     | 229         | 135                   | 17       | 17               |
| Assembly Length         | 1196.17 Mb  | 1196.19 Mb            | 1190.66 Mb | 1190.67 Mb |
| Average Contig Length   | 5.22 Mb     | 8.86 Mb               | 70.04 Mb | 70.04 Mb         |
| Total Length of Ns      | 0           | 18200                 | 8488     |                  |
| N50 value is            | 28.88 Mb    | 95.88 Mb              | 95.88 Mb | 95.88 Mb         |
| N90 value is            | 7.58 Mb     | 76.48 Mb              | 76.48 Mb | 76.29 Mb         |

| **Table 2** BUSCO and LAI scores for the *G. longicalyx* genome compared to existing cotton genomes |
|-------------------------------------------------|-----------------|-----------------|-----------------|-----------------|
| **G. longicalyx**                              | Complete BUSCO | Incomplete BUSCO | LAI score       | Reference       |
|                                                 | Total | Single | Duplicated | Fragmented | Missing |                   |                         |
| G. turneri                                      | 95.80% | 86.50% | 9.30%      | 1.40%      | 2.80%   | 10.74          | (Udall et al. 2019) |
| G. raimondii (BYU)                              | 95.80% | 86.00% | 9.80%      | 1.00%      | 3.20%   | 8.51           | (Udall et al. 2019) |
| G. raimondii (UGI)                              | 92.80% | 85.10% | 7.70%      | 2.70%      | 4.50%   | 10.57          | (Paterson et al. 2012) |
| G. arboreum (CRI)                               | 98.00% | 87.30% | 10.70%     | 0.70%      | 1.30%   | 8.51           | (Du et al. 2018) |
| G. barbadense 3-79 (HAU v2)                     | 84.50% | 12.20% | 84.10%     | 0.80%      | 2.90%   | 10.38          | (Wang et al. 2019) |
| G. hirsutum TM1 (HAU v1)                        | 97.70% | 14.50% | 83.20%     | 0.50%      | 1.80%   | 10.61          | (Wang et al. 2019) |

---

* Statistics for *Longicalyx_V2.0* not calculated.
** Genome size for *G. longicalyx* is 1311 Mb (Hendrix and Stewart 2005).
Genes in species-specific clade, whose species are slightly bigger in total size and have slightly longer sequences, somewhat less than the closely related sister (A-genome) G. herbaceum. Approximately half of the genome (660 Mb) is composed of repetitive elements, as expected from its intermediate genome size (1311 Mb; G. arboreum) vs. G. herbaceum (1667 Mb; G. arboreum). The repetitive fraction is composed of Ty3/gypsy elements, a similar proportion to the proportion of Ty3/gypsy in subgenus Gossypium genomes. Most other element categories were roughly similar in total amount and proportion between G. longicalyx and the two subgenus Gossypium species (Supplemental Figure 2).

Chromatin accessibility in G. longicalyx

We performed ATAC-seq to map accessible chromatin regions (ACRs) in leaves. Two replicated ATAC-seq libraries were sequenced to ~25.7 Gb, and ~45.0 million reads per sample. The strand cross-correlation statistics supported the high quality of the ATAC-seq data, and the correlation of mapping read coverages (Pearson $r = 0.99$ and Spearman $r = 0.77$) suggested a high level of reproducibility between replicates (Supplemental Table 2). A total of 28,030 ACRs (6.4 Mb) were identified ranging mostly from 130 bp to 400 bp in length, which corresponds to ~0.5% of the assembled genome size (Supplemental Table 3). The enrichment of ACRs around gene transcription start sites (Supplemental Figure 3) suggested that these regions were functionally important and likely enriched with cis-regulatory elements. Based on proximity to their nearest annotated genes, these ACRs were categorized as genomic (gACRs; overlapping a gene), proximal (pACRs; within 2 Kb of a gene) or distal (dACRs; >2 Kb from a gene). The gACRs and pACRs represented 12.2% and 13.2% of the total number of ACRs (952 Kb and 854 Kb in size, respectively), while approximately 75% (4.6 Mb) were categorized as dACRs, a majority of which were located over 30 Kb from the nearest gene (Figure 1). This high percentage of dACRs is greater than expected (~40% of 1 GB genome) given previous ATAC-seq studies in plants (Lu et al. 2019; Ricci et al. 2019) and may reflect challenges in annotating rare transcripts. While more thorough, species-specific RNA-seq will improve later annotation versions and refine our understanding of ACR proximity to genes, we do note that our observation of abundant dACRs and potentially long-range cis-regulatory elements is consistent with previous results (Lu et al. 2019; Ricci et al. 2019) The dACRs discovered here were the most GC-rich, followed by gACRs and pACRs (52%, 46%, and 44%, respectively), all of whom had GC contents significantly higher than randomly selected control regions with the same length distribution (Figure 1d). Because high GC content is associated with several distinct features that can affect the cis-regulatory potential of a sequence (Landolin et al. 2010; Wang et al. 2012), these results support the putative regulatory functions of ACRs.

**Genomics of G. longicalyx reniform nematode resistance**

Reniform nematode is an important cotton parasite that results in stunted growth, delayed flowering and/or fruiting, and a reduction in both yield quantity and quality (Robinson 2007; Khanal et al. 2018). While domesticated cotton varieties are largely vulnerable to reniform nematode (Robinson et al. 1997), nematode resistance is found in some wild relatives of domesticated cotton, including G. longicalyx, which is nearly immune (Yik and Birchfield 1984). Recent efforts to elucidate the genetic underpinnings of this resistance...
in *G. longicalyx* (i.e., RENlon) identified a marker (BNL1231) that consistently cosegregates with resistance and is flanked by the SNP markers GI_168758 and Gl_072641 (Dighe et al. 2009; Zheng et al. 2016). Located in chromosome 11, this region contains one or more closely-linked nearly dominant gene(s) (Dighe et al. 2009) that confer hypersensitivity to reniform infection (Khanal et al. 2018), resulting in the “stunting” phenotype; however, the possible effects of co-inherited R-genes has not been eliminated. Because the introgressed segment recombines at reduced rates in interspecific crosses, it has been difficult to fine-map the gene(s) of interest. Additionally, progress from marker-assisted selection has been lacking, as no recombinants have possessed the desired combination of reniform resistance and “non-stunting” (Zheng et al. 2016). Therefore, more refined knowledge of the position, identity of the resistance gene(s), mode(s) of immunity, and possible causes of “stunting” will likely catalyze progress on nematode resistance.

BLAST analysis of the three RENlon-associated markers (above) to the assembled *G. longicalyx* genome identifies an 850 kb region on chromosome F11 (positions 94747040..95596585; Figure 2) containing 52 predicted genes (Supplemental Table 4). Functional annotation reveals that over half of the genes (29, or 56%) are annotated as “TMV resistance protein N-like” or similar. In tobacco, TMV resistance protein N confers a hypersensitive response to the presence of the tobacco mosaic virus (TMV; (Erickson et al. 1999). Homologs of this gene in different species can confer resistance to myriad other parasites and pathogens, including aphid and nematode resistance in tomato (Rossi et al. 1998); fungal resistance in potato (Hehl et al. 1999) and flax (Ellis et al. 2007); and viral resistance in pepper (Guo et al. 2017). Also included in this region are 6 genes annotated as strictosidine synthase-like (SSL), which may also function in immunity and defense (Sohani et al. 2009). While the six SSL-like genes are tandemly arrayed without disruption, several other genes are intercalated within the array of TMV resistance-like genes, including the 6 SSL-like genes (Supplemental Table 4).

Because there is agronomic interest in transferring nematode resistance from *G. longicalyx* to other species, we generated orthogroups between *G. longicalyx*, the two domesticated polyploid species (i.e., *G. hirsutum* and *G. barbadense*), and their model diploid progenitors (*G. raimondii* and *G. arboreum*; Supplemental Table 5; Supplemental File 2). Interestingly, many of the defense-relevant *G. longicalyx* genes in the RENlon region did not cluster into orthogroups with any other species (15 out of 38; Table 5), including 11 of the 29 TMV resistance-related genes in the RENlon region, and fewer were found in syntenic positions in *G. arboreum*. Most of the TMV resistance-related genes that cluster between *G. longicalyx* and other *Gossypium* species are present in a single, large orthogroup (OG0000022; Table 5), whereas the remaining TMV-resistance like genes from *G. longicalyx* are commonly in single gene orthogroups. Since disease resistance (R) proteins operate by detecting specific molecules elicited by the pathogen during infection (Martin et al. 2003), the increased copy number and variability among the *G. longicalyx* TMV-resistance-like genes may suggest specialization among copies.

**Comparative genomics and the evolution of spinnable fiber**

Cotton fiber morphology changed dramatically between *G. longicalyx* and its sister clade, composed of the A-genome cottons *G. arboreum*
Table 5 Orthogroup identity (by Orthofinder) for defense-related genes in the Ren<sup>1000</sup> region and the copy number per species. In G. longicalyx, this number includes genes found outside of the Ren<sup>1000</sup> region. G. hirsutum and G. barbadense copy numbers are split as genes found on the A or D chromosomes, or on scaffolds/contigs not placed on a chromosome.

| Description | Orthogroup | G. longicalyx gene in Ren<sup>1000</sup> region | G. longicalyx | G. arboreum | G. raimondii | G. hirsutum | G. barbadense |
|-------------|------------|-----------------------------------------------|---------------|-------------|--------------|--------------|--------------|
| adenyl-sulfate kinase 3-like | OG0053444 | Golon.011G359300 | 1 | — | — | — | — |
| L-type lectin-domain containing receptor kinase IV.2-like | OG0053450 | Golon.011G361200 | 1 | — | — | — | — |
| T-complex protein 1 subunit theta-like | OG0053447 | Golon.011G360400 | 1 | — | — | — | — |
| protein STRICOSIDINE SYNTHASE-LIKE 10-like | OG0000242 | Golon.011G363400 | 6 | 4 | 2 | 6 A | 9 A, 5 scaffold |
| TMV resistance protein N-like | OG0053454 | Golon.011G361000 | 1 | — | — | — | — |
| TMV resistance protein N-like isoform X1 | OG0053449 | Golon.011G360900 | 1 | — | — | — | — |
| TMV resistance protein N-like isoform X2 | OG0028874 | Golon.011G359900 | 4 | — | — | — | — |

*This gene is syntenically conserved with G. arboreum in the COGE-GEVO analysis.

**This orthogroup is split between two related, but separately named, annotations.*
Figure 3  Synteny between G. longicalyx and domesticated G. arboreum. Mean percent identity is illustrated by the color (93–94% identity from blue to red), including intergenic regions. Lower right inset: Distribution of pairwise p-distances between coding regions of predicted orthologs (i.e., exons only, start to stop) between G. longicalyx and either G. arboreum (blue) or G. raimondii (green). Only orthologs with <5% divergence are shown, which comprises most orthologs in each comparison.
and G. herbaceum. Whereas G. longicalyx fibers are short and tightly adherent to the seed, A-genome fibers are longer and suitable for spinnning. Accordingly, there has been interest in the changes in the A-genome lineage that have led to spinnable fiber (Hovav et al. 2008; Paterson et al. 2012). Progress here has been limited by the available resources for G. longicalyx, relying on introgressive breeding (Nacoulima et al. 2012), microarray expression characterization (Hovav et al. 2008), and SNP-based surveys (Paterson et al. 2012) of G. longicalyx genes relative to G. herbaceum. As genomic resources and surveys for selection are becoming broadly available for the A-genome cottons, our understanding of the evolution of spinnable fiber becomes more tangible by the inclusion of G. longicalyx.

Whole-genome alignment between G. longicalyx and the closely related G. arboreum (domesticated for long fiber) shows high levels of synteny and overall sequence identity (Figure 3). In general, these two genomes are largely collinear, save for scattered rearrangements and several involving chromosomes 1 and 2; these latter may represent a combination of chromosomal evolution and/or misassembly in one or both genomes. Notably, comparison of G. longicalyx to other recently published genomes (Supplemental Figures 4–7) suggests that an inversion in the middle of G. arboreum Chr01 exists relative to representatives of the rest of the genus; however, the other structural rearrangements are restricted to G. arboreum and its derived A subgenome in G. hirsutum and G. barbadense, suggesting that these differences are limited to comparisons between G. longicalyx and A-(sub)genomes.

Genic comparisons between G. longicalyx and G. arboreum suggests a high level of conservation. Orthogroup analysis finds a one-to-one relationship between these two species for over 70% of genes. Most of these putative orthologs exhibit <5% divergence (p-distance) in the coding regions, with over 50% of all putative orthologs exhibiting less than 1.5% divergence. Comparatively, the median divergence for putative orthologs between G. longicalyx and the more distantly related G. raimondii is approximately 2%, with ortholog divergence generally being higher in the G. raimondii comparison (Figure 3, inset).

Because G. longicalyx represents the ancestor to spinnable fiber, orthogroups containing only G. arboreum or polyploid A-genome gene annotations may represent genes important in fiber evolution. Accordingly, we extracted 705 G. arboreum genes from orthogroups composed solely of G. arboreum or polyploid (i.e., G. hirsutum or G. barbadense) A-genome gene annotations for BLAST and functional annotation. Of these 705 genes, only 20 represent genes known to influence fiber, i.e., ethylene responsive genes (10), auxin responsive genes (5), and peroxidase-related genes (5 genes; Supplemental Table 5). While other genes on this list may also influence the evolution of spinnable fiber, identifying other candidates will require further study involving comparative coexpression network analysis or explicit functional studies.

CONCLUSION

While several high-quality genome sequences are available for both wild and domesticated cotton species, each new species provides additional resources to improve both our understanding of evolution and our ability to manipulate traits within various species. In this report, we present the first de novo genome sequence for G. longicalyx, a relative of cultivated cotton. This genome not only represents the ancestor to spinnable fiber, but also contains the agronomically desirable trait of reniform nematode immunity. This resource forms a new foundation for understanding the source and mode of action that provides G. longicalyx with this valuable trait, and will facilitate efforts in understanding and exploiting it in modern crop species.

ACKNOWLEDGMENTS

We thank Emma Miller and Evan Long for technical assistance. We thank the National Science Foundation Plant Genome Research Program (Grant #1339412) and Cotton Inc. for their financial support. This research was funded, in part, through USDA ARS Agreements 58-6066-6-046 and 58-6066-6-059. Support for R.J.S and Z.L. was provided by NSF IOS-1856627 and the Pew Charitable Trusts. We thank BYU Fulton SuperComputer lab for their resources and generous support. We also thank ResearchIT for computational support at Iowa State University. We thank Rise Services for office accommodations in Orem, UT.

LITERATURE CITED

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. 215: 403–410. https://doi.org/10.1016/0022-2836(82)90546-2
Bailly-Bechet, M., A. Haudry, and E. Lerat, 2014 “One code to find them all”: a perl tool to conveniently parse RepeatMasker output files. Mob. DNA 5: 13. https://doi.org/10.1186/1759-8753-5-13
Bao, W., K. K. Koijima, and O. Kohany, 2015 Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA 6: 11. https://doi.org/10.1186/s13100-015-0041-9
Bell, A., and A. F. Robinson, 2004 Development and characteristics of triple species hybrids used to transfer reniform nematode resistance from Gossypium longicalyx to Gossypium hirsutum, pp. 422–426 in Proceedings of the Beltwide Cotton Conferences, naldc.nal.usda.gov.
Birchfield, W., and L. R. Brister, 1963 Susceptibility of cotton and relatives to reniform nematode in Louisiana. Plant Dis. Rep. 47: 990–992.
Boetzer, M., and W. Pirovano, 2012 Toward almost closed genomes with GapFiller. Genome Biol. 13: R56. https://doi.org/10.1186/gb-2012-13-6-r56
Borendsova, M., and A. Lomsadze, 2011 Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. Curr. Protoc. Bioinformatics Chapter 4: Unit 4.6.1–Unit 4.6.10. https://doi.org/10.1002/0471250953.bi0406s35
Buenrostro, J. D., P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf, 2013 Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10: 1213–1218. https://doi.org/10.1038/nmeth.2688
Campillo, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos et al., 2009 BLAST+: architecture and applications. BMC Bioinformatics 10: 421. https://doi.org/10.1186/1471-2105-10-421
Campbell, M. S., C. Holt, B. Moore, and M. Yandell, 2014 Genome Annotation and Curation Using MAKER and MAKER-P. Curr. Protoc. Bioinformatics. 421.https://doi.org/10.1002/0471250953.bi0411s48
Chen, Z., K. Feng, C. E. Grover, P. Li, F. Liu et al., 2016 Chloroplast DNA Structural Variation, Phylogeny, and Age of Divergence among Diploid Cotton Species. PLoS One 11: e0157183. https://doi.org/10.1371/journal.pone.0157183
Dighe, N. D., A. F. Robinson, A. A. Bell, M. A. Menz, R. G. Cantrell et al., 2009 Linkage Mapping of Resistance to Reniform Nematode in Cotton following Introgression from Gossypium longicalyx (Hutch. & Lee). Crop Sci. 49: 1151–1164. https://doi.org/10.2135/cropsci2008.03.0129
Du, X., G. Huang, S. He, Z. Yang, G. Sun et al., 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
Ellbeck, K., B. Moore, C. Holt, and M. Yandell, 2009 Quantitative measures for the management and comparison of annotated genomes. BMC Bioinformatics 10: 67. https://doi.org/10.1186/1471-2105-10-67
Ellis, J. G., P. N. Dodds, and G. J. Lawrence, 2007 Flax rust resistance gene specificity is based on direct resistance-avirulence protein interactions. Annu. Rev. Phytopathol. 45: 289–306. https://doi.org/10.1146/annurev.phyto.45.062806.094331
