G3, 2021, 11(11), jkab319
DOI: 10.1093/g3journal/jkab319
Advance Access Publication Date: 6 September 2021
Genome Report

The *Gossypium anomalum* genome as a resource for cotton improvement and evolutionary analysis of hybrid incompatibility

Corrinne E. Grover ,[1] Daojun Yuan,[2] Mark A. Arick II,[3] Emma R. Miller,[1] Guanjing Hu ,[4,5,6] Daniel G. Peterson,[3] Jonathan F. Wendel ,[1] and Joshua A. Udall [7,*]

[1]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50010, USA,
[2]College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, Hubei 430070, China,
[3]Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, MS 39762, USA,
[4]State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences, Anyang 455000, China,
[5]Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture,
[6]Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China, and
[7]USDA/Agricultural Research Service, Crop Germplasm Research Unit, College Station, TX 77845, USA

*Corresponding author: Crop Germplasm Research Unit, USDA-ARS, 2881 F&B Road, College Station, TX 77845, USA. Email: joshua.udall@usda.gov

Abstract

Cotton is an important crop that has been the beneficiary of multiple genome sequencing efforts, including diverse representatives of wild species for germplasm development. *Gossypium anomalum* is a wild African diploid species that harbors stress-resistance and fiber-related traits with potential application to modern breeding efforts. In addition, this species is a natural source of cytoplasmic male sterility and a resource for understanding hybrid lethality in the genus. Here, we report a high-quality *de novo* genome assembly for *G. anomalum* and characterize this genome relative to existing genome sequences in cotton. In addition, we use the synthetic allopolyploids 2(A2D1) and 2(A2D3) to discover regions in the *G. anomalum* genome potentially involved in hybrid lethality, a possibility enabled by introgression of regions homologous to the D3 (*Gossypium davidsonii*) lethality loci into the synthetic 2(A2D3) allopolyploid.

Keywords: *Gossypium anomalum*; genome sequence; PacBio

Introduction

The genus *Gossypium* is responsible for providing a majority of natural textile fiber through cultivation of its four independently domesticated species. Recent efforts in genome sequencing have resulted in high-quality genomes for all domesticated species (Yuan *et al.* 2015; Chen *et al.* 2020; Huang *et al.* 2020) and for other important species (Paterson *et al.* 2012; Udall *et al.* 2019; Chen *et al.* 2020). Recent efforts at sequencing additional wild cotton species (Udall *et al.* 2019; Grover *et al.* 2020, 2021) have resulted in several high-quality resources for exploring the evolution of agronomically favorable traits, *e.g.* stress resistance, that are found naturally in the wild cotton species.

Comprising more than 50 known species, the diploid species of cotton have been placed into genome groups (known as A–G and K) based on meiotic chromosome associations and sequence similarities (see Wang *et al.* 2018 for review). The wild African species *Gossypium anomalum* Waw. & Peyr. is one of the four species comprising the "B-genome" cottons. The B-genome cottons are *G. anomalum* (B1), *Gossypium triphyllum* (B2), *Gossypium capitis-viridis* (B3), and perhaps the poorly understood *Gossypium trifurcatum* (Vollesen 1987; Fryxell 1992; Wendel *et al.* 2010), although relationships for the latter, rare species are unclear (Wang *et al.*, 2018). All of these species are in clades that are close relatives of the diploid domesticated species *Gossypium arboreum* and *Gossypium herbaceum*. *Gossypium anomalum* has a large but disjunct geographic range, encompassing southwest Africa, centered in Namibia (*G. anomalum* subsp. *anomalum*), and then also a broad distribution in northern Africa (*G. anomalum* subsp. *senarense*) (Vollesen 1987; Fryxell 1992). Although the species has no obvious traits of agronomic interest, *G. anomalum* has many understudied characteristics that may be useful in breeding programs and understanding the evolution of favorable phenotypes. The fiber of *G. anomalum* is short and not spinnable, but *G. anomalum* has been considered a potential source for fiber fineness and strength (Mehetre 2010), and the xerophytic nature of *G. anomalum* makes it a candidate for understanding drought resistance in cotton species. *Gossypium anomalum* also exhibits natural resistance to various cotton pests, including jassids (Mammadov *et al.* 2018), bacterial blight/blackarm (Knight 1954; Fryxell 1984), mites (Mehetre 2010), bollworms (Mehetre 2010), and rust (Fryxell 1984; Mehetre 2010). Mechanisms underlying pest resistance are understudied, but it is clear that investigation of the *G. anomalum* genome may illuminate valuable genes and alleles underlying resistance (Fryxell 1984), as demonstrated by hybridization

Received: June 15, 2021. Accepted: August 30, 2021

experiments involving *G. anomalum* and cultivated cottons (Mehetre 2010).

In addition to stress resistance and fiber quality traits, *G. anomalum* is both a source of cytoplasmic male sterility (Meyer and Meyer 1965; Marshall *et al.* 1974) and one of the few cotton species that can be crossed with cottons from subsection *Integrifolia* [*i.e. Gossypium klotzschianum* and *G. davidsonii* (Hutchinson *et al.* 1947)], which generally exhibit hybrid lethality in other crosses. Both cytoplasmic male sterility and *Integrifolia* derived lethality have applications in cotton (Weaver and Weaver 1977; Lee 1981a; Stelly *et al.* 1988; Stelly 1990; Suzuki *et al.* 2013; Bohra *et al.* 2016), the latter being accessible only in crosses that are non-lethal, *e.g.* with *G. anomalum*.

Here we describe a high-quality, *de novo* genome assembly for *G. anomalum*, the first for a member of *Gossypium* section *Anomala*, which are colloquially known as the "B-genome" cottons (Wang *et al.* 2018). This genome provides a genetic repository for investigating potentially valuable agronomic traits.

Methods and materials

Plant material and sequencing methods

Gossypium anomalum was grown from seed under greenhouse conditions at Brigham Young University (BYU), and mature leaves were collected for sequencing. High-quality DNA was extracted via CTAB (Kidwell and Osborn 1992) and subsequently quantified using a Qubit Fluorometer (ThermoFisher, Inc.). DNA was size selected for fragments >18 kb on the BluePippen (Sage Science, LLC) prior to library construction; fragment size was verified using a Fragment Analyzer (Advanced Analytical Technologies, Inc). A single Pacific Biosciences (PacBio) sequencing library was constructed by the BYU DNA Sequencing Center (DNASC), and 15 PacBio cells were sequenced using the Sequel system. Raw reads were assembled using Canu V1.4 with default parameters (Koren *et al.* 2017).

Leaf tissue was shipped to PhaseGenomics LLC for DNA extraction and HiC library construction. HiC libraries were sequenced on the Illumina HiSeq 2500 (PE125 bp) at the BYU DNASC, and the resulting reads were used to join contigs. JuiceBox (Durand *et al.* 2016) was used in conjunction with the HiC reads to correct the assembly based on the association frequency between paired-ends. A custom python script (available through PhaseGenomics, LLC) was used to construct the final genome sequence of *G. anomalum*, which consists of 13 scaffolds corresponding to the haploid complement of chromosomes.

Repeat and gene annotation

Repeats were identified using RepeatMasker (Smit *et al.* 2015) and a custom library consisting of Repbase 23.04 repeats (Bao *et al.* 2015) with cotton-specific repeats (Grover *et al.* 2020). RepeatMasker run parameters were set to a high-sensitivity scan that only masked transposable elements (TEs). Multiple hits were aggregated using "One code to find them all" using default parameters (Bailly-Bechet *et al.* 2014), and the resulting output was summarized in R/4.0.3 (R Core Team 2020) using *dplyr*/2.0.0 (Wickham *et al.* 2015). Repeats were quantified relative to other cotton species with resequencing data downloaded from the GenBank Short-Read Archive (Supplementary Table S1) using the RepeatExplorer pipeline (Novák *et al.* 2010), and results were parsed in R/4.0.3 (R Core Team 2020) as previously described (Grover *et al.* 2020). All codes are available at https://github.com/Wendellab/anomalum (last accessed 9/7/21).

Genome annotations were conducted using existing RNA-seq data from tissues of closely related species (Supplementary Table S2) as previously described (Grover *et al.* 2021). Hisat2 was used to map each RNA-seq library to the hard-masked *G. anomalum* genome (v2.1.0) (Kim *et al.* 2015), and *de novo* transcriptome assemblies were generated via StringTie (v2.1.1) (Pertea *et al.* 2015) and Cufflinks (v2.2.1) (Ghosh and Chan 2016). These RNA-seq assemblies were combined with a Trinity (v2.8.6) (Grabherr *et al.* 2011) reference-guided transcriptome assembly and splice junction information from Portcullis (v1.2.2) (Mapleson *et al.* 2018) in Mikado (v1.2.4) (Venturini *et al.* 2018). GeneMark (v4.38) (Borodovsky and Lomsadze 2011) generated annotations were used in BRAKER2 (v2.1.2) (Hoff *et al.* 2019) to train Augustus (v3.3.2) (Stanke *et al.* 2006). MAKER2 (v2.31.10) (Holt and Yandell 2011; Campbell *et al.* 2014) integrated gene predictions from all three sources, *i.e.* BRAKER2 trained Augustus, GeneMark, and Mikado, with additional evidence from all available *Gossypium* ESTs (NCBI nt database with the filters "txid3633" and "is_est"), all curated proteins in Uniprot Swissprot (v2019_07) (UniProt Consortium 2008), and all annotated proteins from the *Gossypium hirsutum* (https://www.cottongen.org/species/Gossypium_hirsutum/jgi-AD1_genome_v1.1; last accessed 9/7/21) and *Gossypium raimondii* (Paterson *et al.* 2012) genomes.

Each gene model was scored by Maker using the annotation edit distance [AED; (Eilbeck *et al.* 2009; Holt and Yandell 2011; Yandell and Ence 2012)] relative to EST and protein evidence, and gene models with an AED < 0.37 were retained. These gene models were functionally annotated using InterProScan (v5.47-82.0) (Jones *et al.* 2014) and BlastP (v2.9.0+) (Camacho *et al.* 2009) searches against the Uniprot SwissProt database. Orthologous relationships between *G. anomalum* and other sequenced diploid cotton genomes, *i.e. Gossypium longicalyx* (Grover *et al.* 2020), *G. arboreum* (Du *et al.* 2018), *G. herbaceum* (Huang *et al.* 2020), and *G. raimondii* (Paterson *et al.* 2012) are derived from previously published (Grover *et al.* 2021) Orthofinder analyses (Emms and Kelly 2015, 2019). All genomes are hosted through CottonGen (https://www.cottongen.org; Yu *et al.* 2014; last accessed 9/7/21)) and running parameters are available from https://github.com/Wendellab/anomalum.

Gossypium anomalum introgression in the synthetic allotetraploid, 2(A2D3)

A synthetic allotetraploid, *i.e.* 2(A2D3), was generated by Joshua Lee in the late 1970s to early 1980s. The first step in producing the allotetraploid involved crossing *G. anomalum* (B1) with *G. arboreum* (A2; Supplementary Figure S1). The latter species is incompatible with *G. davidsonii* (D3), as are all species tested except *G. anomalum*; *G. anomalum* likely possesses a null allele for the lethality locus. By repeatedly backcrossing the *G. anomalum* compatibility region into the recipient *G. arboreum* parent, Lee was able to create a *G. arboreum* line that was compatible with *G. davidsonii*. This was subsequently used to generate a diploid hybrid with *G. davidsonii*, *i.e.* A2 × D3. Subsequent doubling of this hybrid generated the synthetic 2(A2D3), and this plant has been subsequently maintained by Jonathan Wendel in the Iowa State University greenhouse since the mid-1980s. We downloaded previously generated reads from this synthetic allotetraploid (Supplementary Table S1), along with reads for an additional synthetic allotetraploid, 2(A2D1) (Beasley 1940). Chromosomes from all three diploid species, *i.e. G. arboreum*, *G. anomalum*, and *G. davidsonii*, were combined to generate an *in silico* genome designated "ABD". Mapping of the 2(A2D3) reads to the ABD genome identified reads which best match *G. anomalum*. To verify the

mapping results, we also mapped reads from *G. arboreum*, *G. davidsonii*, and the 2(A2D1) synthetic to the same (synthetic) ABD genome. The 2(A2D1) synthetic was included as an additional control because the *G. arboreum* (A2) used in this initial cross [*i.e.*, *var. neglectum* (Beasley 1940)] did not include known introgression from other cotton species. All reads were mapped to the ABD genome using BWA (v0.7.17) (Li and Durbin 2009), and samtools (v1.9) (Li *et al.* 2009) was used to select the reads from each species that uniquely mapped (mapq $\geq$ 30) to the *G. anomalum* genome. Contiguous regions of uniquely mapped reads were combined in bedtools (v2.28.0) (Quinlan 2014) for each of the control libraries, *i.e. G. arboreum*, *G. davidsonii*, and 2(A2D1) to identify putative regions of ambiguity (*i.e.* where reads may preferentially map to the *G. anomalum* chromosomes by chance). Overlapping regions between the mapping results of 2(A2D1) and 2(A2D3) were filtered to retain regions where only 2(A2D3) reads mapped to the *G. anomalum* genome sequence. Regions < 5 kb in length were then discarded. These filters resulted in a high-confidence set of reads that were likely derived from the *G. anomalum* introgression specific to 2(A2D3).

Results and discussion

Genome assembly and annotation

Here, we report a *de novo* genome assembly for *G. anomalum* using 55× coverage of PacBio reads and 140.5 million (M) HiC reads. The initial assembly yielded 229 contigs with a N50 of 11 Mb. HiC information was integrated to produce a more contiguous assembly, consisting of 13 chromosomes with an average length of 92 Mb and containing only 20.8 kb (0.002%) gap sequence within the chromosomal scaffolds. The total assembly length is 88% of the estimated 1359 Mb genome (Hendrix and Stewart 2005). Full statistics are in Supplementary Table S3.

BUSCO analysis (Waterhouse *et al.* 2017) of the 13 assembled chromosomes recovered 97.1% complete BUSCOs from the 2326 BUSCO groups comprising euidcots_odb10 database, similar to other published cotton diploid genomes (Table 1). In general, most BUSCOs were both complete and single copy (89.5%), with a low level of duplication (7.6%). Few BUSCOs were fragmented (0.5%) or missing (2.4%), which indicates a general completeness of the assembly. Dotplot comparisons to other cotton genomes (Figure 1) reveal high colinearity, as expected from previous comparisons among cotton species (Udall *et al.* 2019; Grover *et al.* 2020, 2021). While inversions and other small structural differences are apparent between *G. anomalum* and other published genomes (Figure 1), these differences are restricted to one or a few genomes and likely reflect lineage-specific differences rather than misassembly. This is supported by Hi-C mapping (Supplementary Figure S2), which both supports the structure of the assembled genome and further confirms that the *G. anomalum* assembly is similar to or superior to recently published genomes.

Genome annotation of the 13 chromosomes produced 37,830 primary transcripts, which is similar to other cotton diploids (Paterson *et al.* 2012; Du *et al.* 2018; Udall *et al.* 2019; Grover *et al.* 2020, 2021; Huang *et al.* 2020; Wang *et al.* 2021) whose gene numbers range between 34,928 (Grover *et al.* 2021) in *Gossypium stocksii* and 43,952 (Huang *et al.* 2020) in *G. herbaceum*. Of these 37,830 predicted genes, 36,802 were functionally annotated via BLAST; 21,768 via GO; 34,916 via InterPro; 29,248 via PFAM; 2897 were annotated via TIGRFAM; 34,916 were annotated by at least two; and 29,280 were annotated by three or more. The *G. anomalum* annotation averaged 5.8 exons and 4.8 introns per gene, whose average lengths (257 and 339 bp, respectively) were similar to previously published cotton species (Table 2). BUSCO analysis of the transcriptome exhibited similar quality to the genome, with 86.1% complete and single copy and few duplicated or missing (8 and 4.8%, respectively). Ortholog analysis of primary transcripts suggests that the pattern of orthogroups including *G. anomalum* is similar to other diploid cotton species, although the number of genes not assigned to orthogroups is fewer than previously noted (Grover *et al.* 2021), whereas the number of species-specific orthogroups is higher, albeit still low (Table 2).

Repeats

Both *de novo* TE prediction (Bailly-Bechet *et al.* 2014; Smit *et al.* 2015) and repetitive clustering (Novák *et al.* 2010) were used to assess repetitive elements in the *G. anomalum* genome. As with *G. longicalyx* (Grover *et al.* 2020), RepeatExplorer estimated a larger proportion of the *G. anomalum* genome as repetitive (46.5%) compared with RepeatMasker (42%). Estimates for the different TE categories surveyed (*e.g.* DNA, Ty3/*gypsy*, Ty1/*copia*, etc.) were generally consistent between the two methods (Supplementary Table S4), although RepeatMasker recovered far more *copia* elements than did RepeatExplorer (47.7 Mbp, *vs* 29.1 Mbp). This is likely due to the inability of RepeatExplorer to efficiently categorize *copia*-like elements in this genome, instead placing them in a general "LTR" category (21.9 Mbp, *vs* 0 Mbp for RepeatMasker). As is common among plants, most of the repetitive sequence recovered by both methods was attributed to *gypsy* elements, which occupy 38% of the genome according to RepeatMasker and 42% of the genome based on RepeatExplorer analysis.

In addition to characterizing the *G. anomalum* genome via clustering, we also co-clustered a diverse array of previously sequenced species (see methods) to evaluate the repeat content of *G. anomalum* in the broader context of the genus (Supplementary

Table 1 BUSCO scores for the genome and transcriptome of *G. anomalum*, as compared with genome BUSCO from recent cotton diploids

	Complete BUSCO			Incomplete BUSCO		
	Total	Single	Duplicated	Fragmented	Missing	References[a]
G. anomalum (genome)	97.1%	89.5%	7.6%	0.5%	2.4%	
G. stocksii	97.6%	88.9%	8.7%	0.9%	1.5%	(Grover *et al.* 2021)
G. longicalyx	95.8%	86.5%	9.3%	1.4%	2.8%	(Grover *et al.* 2020)
G. turneri	95.8%	86.0%	9.8%	1.0%	3.2%	(Udall *et al.* 2019)
G. raimondii (BYU)	92.8%	85.1%	7.7%	2.7%	4.5%	(Udall *et al.* 2019)
G. raimondii (JGI)	98.0%	87.3%	10.7%	0.7%	1.3%	(Paterson *et al.* 2012)
G. arboreum (CRI)	94.7%	85.2%	9.5%	1.0%	4.3%	(Du *et al.* 2018)
G. anomalum (annotation)	94.1%	86.1%	8.0%	1.1%	4.8%	

[a] Reference for each genome is given.
BUSCO scores for existing genomes are as reported in (Grover *et al.* 2020, 2021).

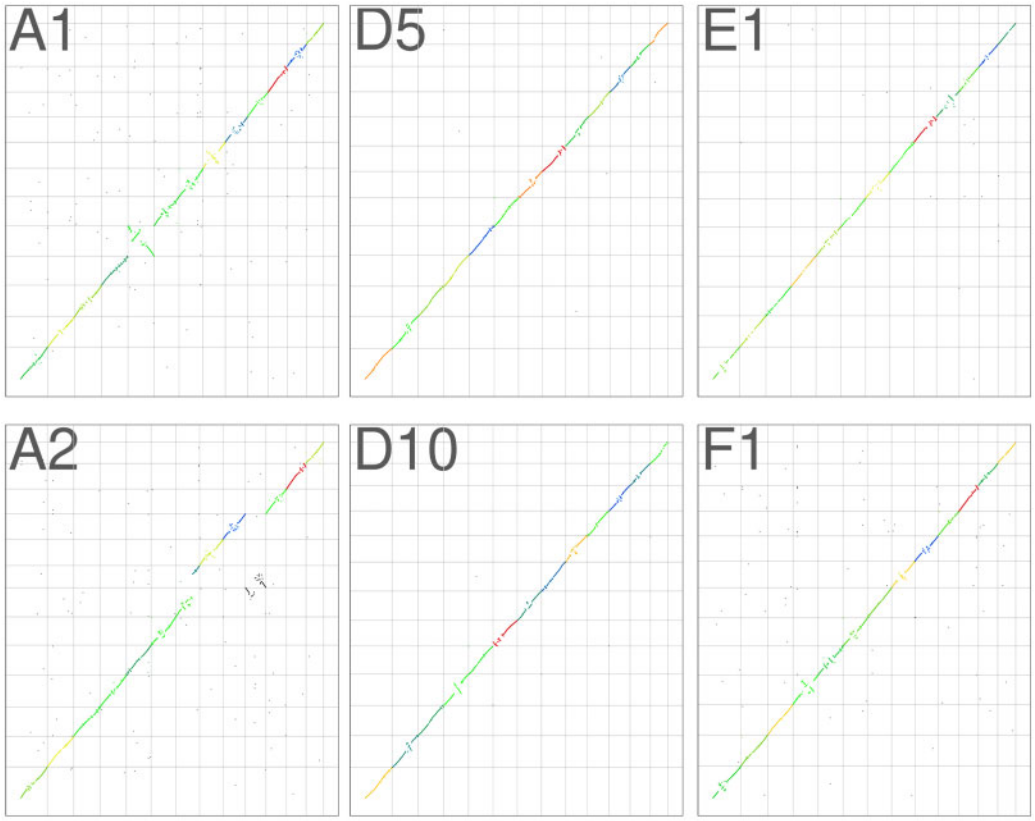

Figure 1 Pairwise comparisons of *G. anomalum* with *G. herbaceum* (A1) (Huang *et al.* 2020), *G. raimondii* (D5) (Udall *et al.* 2019), *G. stocksii* (Grover *et al.* 2021), *G. arboreum* (A2) (Huang *et al.* 2020), *G. turneri* (D10) (Udall *et al.* 2019), and *G. longicalyx* (F1) (Grover *et al.* 2020).

Table S1). This clustering included at least one member of each lettered cotton "genome group" (*i.e.* A–G and K; Wang *et al.* 2018), which were all sampled to represent 1% of their genome size (Hendrix and Stewart 2005). Principal components analysis (PCA; Figure 2) generally separates the species by geography on the first axis, with the American "D-genome" cottons toward the left part of the plot, the Australian species (groups C, G, and K) toward the right, and the African species (genome groups A, B, E, and F) intermediate between those two. Notably, the PCA groupings loosely follow the phylogenetic relationships among the genome groups (Cronn *et al.* 2002).

Relative to other cotton species, *G. anomalum* (B-genome) has an intermediate amount of TEs (Figure 3; Table 3), as expected from its intermediate genome size. Like most of the other cottons, the *G. anomalum* genome consists of approximately half repetitive sequences (627 Mb), most of which (90%) are *gypsy* elements.

Interestingly, while the *G. anomalum* genome is around 200 Mbp smaller than the African E-genome species (represented here by *Gossypium somalense*), cluster analysis suggests that it has about 60 Mbp *more* repetitive sequences, most of which (40 Mb) are annotated as *gypsy* elements (Table 3). Regression analysis suggests that the amount of repetitive sequence observed in the E-genome clade is lower than expected, given the rest of the genome groups (Supplementary Figure S3). This may indicate general degradation and/or divergence in the repeats found in the E-genome clade, possibly indicating the presence of older elements, and/or that prior estimates of genome size are overestimates. The latter hypothesis would be consistent with the high contiguity and quality of our assembly that nevertheless recovered only 88% of the expected genome size.

Also notable is the observation that while the A-genome species (represented by both extant species, *G. herbaceum* and *G. arboreum*) are only ~350 Mbp larger than *G. anomalum*, clustering suggests that they have approximately 1.5× more repetitive sequences, mostly *gypsy* (927 Mbp in A, *vs* 565 Mbp in B). This is, however, within what is expected as genome sizes in *Gossypium* increase (Supplementary Figure S3).

Gossypium anomalum as a vehicle to understand hybrid lethality in cotton

Hybrid lethality is a postzygotic reproductive barrier that results in embryo and/or seedling death in crosses involving incompatible plants, resulting in reduction and/or elimination of gene flow between populations or species (Bomblies and Weigel 2007; Maheshwari and Barbash 2011). While interspecific incompatibilities are common between species from different genome groups in *Gossypium*, interfertility is quite common between species from the same genome group (Hutchinson 1932; Silow 1941; Stephens 1946; Gerstel 1954; Menzel and Brown 1955; Phillips and Merritt 1972; Phillips and Reid 1975; Lee 1981c). For example, crosses are possible for most combinations of the 14 recognized D-genome diploids, insofar as these have been tested. An exception to this generality involves hybrid lethality in crosses that involve species from subsection *Integrifolia* (*i.e. G. davidsonii* and *G. klotzschianum*; D3d and D3k, respectively). These sister species are incompatible with nearly every other species in the genus, with the exception of *G. longicalyx* (F-genome) and *G. anomalum* (Phillips 1963). Notably, in some cases, this lethality can be circumvented by increasing germination and growth temperatures (Phillips 1977), making lethality potentially useful in cultivar development (Lee 1981a).

Table 2 Gene statistics and relationships between *G. anomalum* and other cotton diploid genomes (primary transcripts only)

	G. anomalum	*G. arboreum*		*G. herbaceum*	*G. raimondii*		*G. turneri*	*G. longicalyx*	*G. australe*	*G. stocksii*	*G. rotundifolium*
		Li et al. 2014	Du et al. 2018	Huang et al.2020	Huang et al. 2020	Paterson et al. 2012	Wang et al. 2012	Udall et al. 2019	Udall et al. 2019	Grover et al. 2020	Cai et al. 2020	Grover et al. 2021	Wang et al. 2021
Number of genes	**37,830**	40,134	40,960	43,278	43,952	37,223	40,976	41,030	38,871	38,378	38,281	34,928	39,355
Genes in orthogroups	**36,847 (97%)**	38,605 (96%)	40,565 (99%)	42,599 (98%)	42,955 (98%)	36,774 (99%)	39,829 (97%)	38,317 (93%)	36,501 (94%)	37,016 (97%)	36,164 (95%)	34,012 (97%)	38,511 (98%)
Unassigned genes	**983 (3%)**	1,529 (4%)	395 (1%)	679 (2%)	997 (2%)	449 (1%)	1,147 (3%)	2,713 (7%)	2,370 (6%)	1,362 (4%)	2,117 (6%)	916 (3%)	844 (2%)
Average exon number	**5.8**	4.6	5.1	4.7	4.7	5.0	4.5	5.7	4.8	5.8	5.4	6.3	5.2
Average exon length	**257**	235	233	245	244	203	243	209	223	255	215	226	230
Average intron number	**4.8**	3.6	4.1	3.7	3.7	4.0	3.5	4.7	3.8	4.8	4.4	5.3	4.2
Average intron length	**339**	288	362	337	330	419	306	253	307	361	362	350	448
Orthogroups containing species	**24,591 (64%)**	27,731 (72%)	28,614 (75%)	28,452 (74%)	29,359 (77%)	27,216 (71%)	26,844 (70%)	26,017 (68%)	26,451 (69%)	25,940 (68%)	20,504 (54%)	24,500 (64%)	25,055 (65%)
Species-specific orthogroups	**109**	130	30	86	107	13	309	137	133	136	634	96	313
Genes in species-specific orthogroups	**389 (1%)**	367 (1%)	90 (0%)	412 (1%)	396 (1%)	30 (0%)	1,421 (4%)	516 (1%)	338 (1%0	406 (1%)	2,794 (7%)	509 (2%)	2,167 (6%)

While loci conferring hybrid lethality have been genetically identified through crosses and/or hexaploid bridging (Lee 1981a,c; Endrizzi *et al.* 1985; Stelly 1990; Samora *et al.* 1994; Song *et al.* 2009), the underlying gene(s) controlling the D3 incompatibility are not yet known. In the late 1970s to early 1980s, Joshua A. Lee generated a synthetic 2(A2D3) allotetraploid (Supplemental Figure S2) as described above, using the trick that *G. anomalum* was apparently "null" for the incompatibility factor and thus could be introgressed into A2 for purposes of creating the novel allopolyploid. Using a scheme of repeated backcrossing into *G. arboreum* and testing for fertility with *G. davidsonii*, crosses were continued for an unknown number of generations, until hybrid progeny were uniformly healthy. Thus, the interspecific F1 hybrids were really tri-species constructs, in part, containing an introgressed locus (or loci) from *G. anomalum* that permits crosses with D3 to survive; ostensibly, this locus codes for a lethality factor in wild-type D3. Progeny from the last successful *G. arboreum* (BC) × *G. davidsonii* were subsequently doubled to create the synthetic 2(A2D3). This synthetic allotetraploid is thus primarily composed of *G. arboreum* and *G. davidsonii*, containing only a residual contribution from *G. anomalum*.

At present, the nature of the gene or genes controlling this hybrid lethality are unknown. Previous cytogenetic work on D3-lethality suggests that a single locus in *G. davidsonii* (*i.e.* Le^{dav}) is responsible for lethality (Lee 1981b), and that this may interact with 1–2 loci in other cotton species (Lee 1981b; Stelly 1990). We downloaded resequencing reads from 2(A2D3) and a second synthetic allotetraploid [*i.e.* 2(A2D1)], which is a doubled *G. arboreum* × *Gossypium thurberi* (Beasley 1940), and thus similar to 2(A2D3) but lacking the *G. anomalum* introgression. Competitive mapping of both synthetic allotetraploids to a reference containing the combined genomes of *G. arboreum*, *G. anomalum*, and *G. raimondii* (*i.e.* hereafter ABD-reference) reveals that approximately 1–2% of reads in each synthetic map strictly to *G. anomalum* chromosomes (Table 4), with a slightly higher percentage of reads from 2(A2D3) characterized as B-like (1.97 *vs* 1.66%). The number of reads considered A- or D-like is over an order of magnitude higher for both synthetics. Reads that could not be distinguished as A-, B-, or D-like (due to shared ancestry) were discarded from all samples, retaining approximately 70–75% of the reads. Because symplesiomorphy, autapomorphy, and technical error all have the potential to confound species identification of reads, we filtered locations in the ABD-reference *G. anomalum* chromosomes where we unexpectedly observed mapping of *G. arboreum*, *G. davidsonii*, and/or 2(A2D1)-derived reads, all of which should not have a *G. anomalum* origin. The remaining regions were considered markers for candidate locations where *G. anomalum* introgression remains in the 2(A2D3) synthetic allotetraploid.

After discarding short (<5 kb) regions as putative artifacts, we identified 28 regions on 9 chromosomes with putative introgression (Supplementary Table S5), for a total length of 195.7 kb. Most chromosomes exhibit small, discontiguous regions of putative introgression (<25 kb in length); however, a 287.5 kb window on chromosome B06 contains 13 of the 28 regions that collective comprise 69% (111.5 kb) of the total introgressed *G. anomalum* sequence. Genome annotation in this putative introgressed hotspot reveals only two gene models (*i.e.* B06G223600 and B06G223900) that overlap with the B-like regions, suggesting that one or both of these genes may be important for conferring fertility with *Integrifolia* (D3) species. Although these gene models are near-sequential in the genome, they are separated by over 172 kb of intervening sequence, as well as two additional genes contained within the 287.5 kb window that do not exhibit evidence of

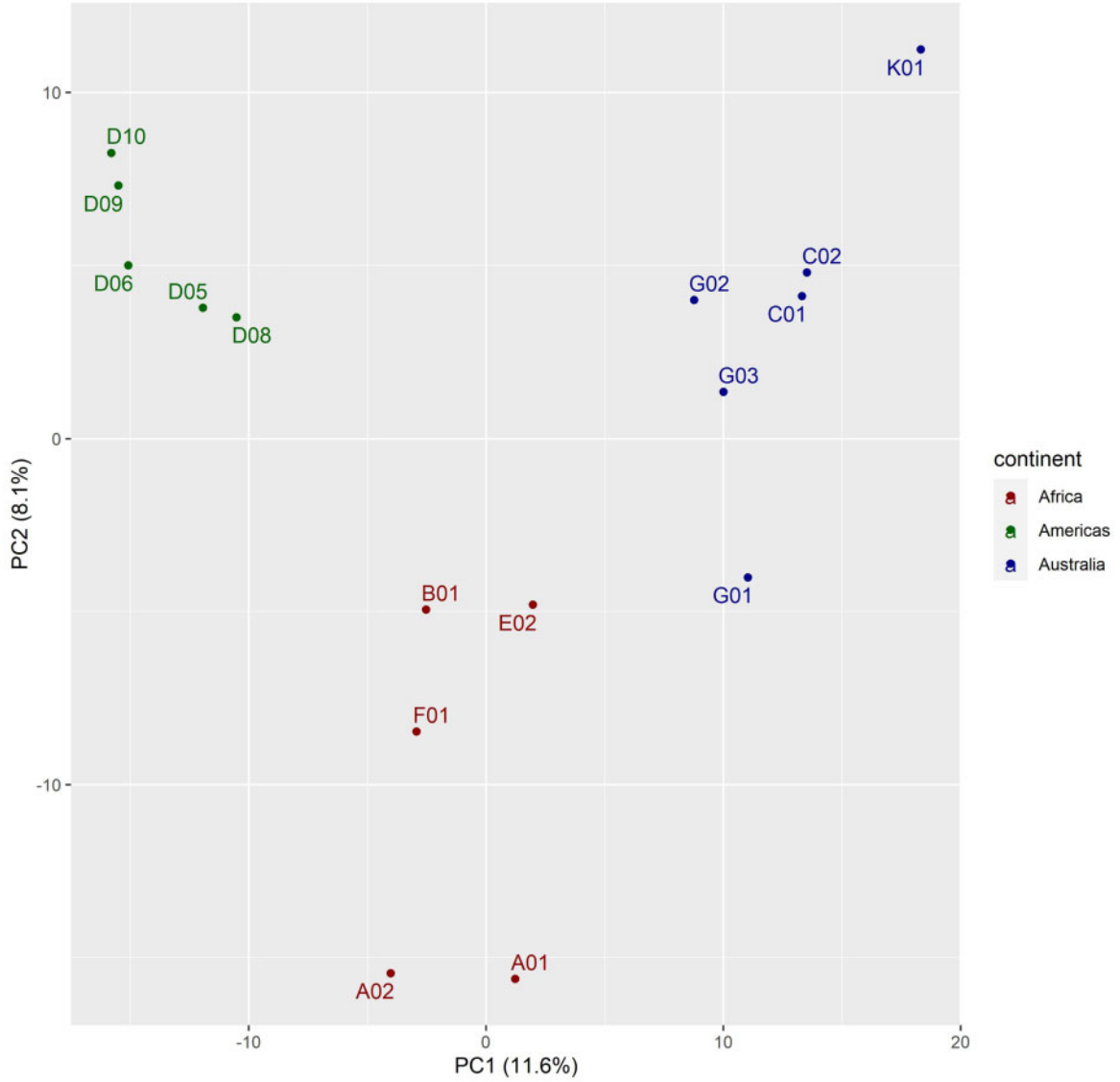

Figure 2 PCA analysis of repeats in cotton species. Species placement on the first two axes is primarily due to a small number of *gypsy* clusters. Species are colored by their broad geographic groups, *i.e.*, the Americas (green), Africa/Arabian Peninsula (red), and Australia (blue) and listed by their official designations (Wang *et al.* 2018). The American cottons are *G. raimondii* (D5), *G. gossypioides* (D6), *G. trilobum* (D8), *G. laxum* (D9), and *G. turneri* (D10). The African/Arabian cottons are *G. herbaceum* (A1), *G. arboreum* (A2), *G. anomalum* (B1), *G. somalense* (E2), and *G. longicalyx* (F1). The Australian cottons are *G. sturtianum* (C1), *G. robinsonii* (C2), *G. bickii* (G1), *G. australe* (G2), *G. nelsonii* (G3), and *G. exiguum* (K1).

Table 3 Repetitive amounts (average) in *Gossypium* genome groups compared with *G. anomalum* (B)

Genome group	Geographic location	Genome size (Mb)	Repeats (Mb)	Repeats (%)	Gypsy (Mb)	Gypsy (% repeats)	Gypsy (% genome)
D	Americas	885	286.6	32%	224	78%	25%
F	Africa	1311	607.4	46%	550	91%	42%
B	**Africa**	**1350**	**627.1**	**47%**	**565**	**90%**	**42%**
E	Africa	1560	567.5	36%	520	92%	33%
A	Africa	1697	992.9	59%	927	93%	55%
G	Australia	1785	1022.0	57%	906	89%	51%
C	Australia	1980	1253.0	63%	1127	90%	57%
K	Australia	2572	1617.6	63%	1465	91%	57%

Species representing each genome group are given in Supplementary Table S1.

introgression. The first gene, B06G223600, is a putative F-box/kelch-repeat protein similar to At4g19870, whereas the second (B06G223900) is similar to *PAP12*, a phosphatase from *Arabidopsis thaliana*. Notably, in 2(A2D3), B06G223600 has 15 bp of extra sequence relative to the A-genome ortholog, representing an additional five amino acids in the protein. The second gene,

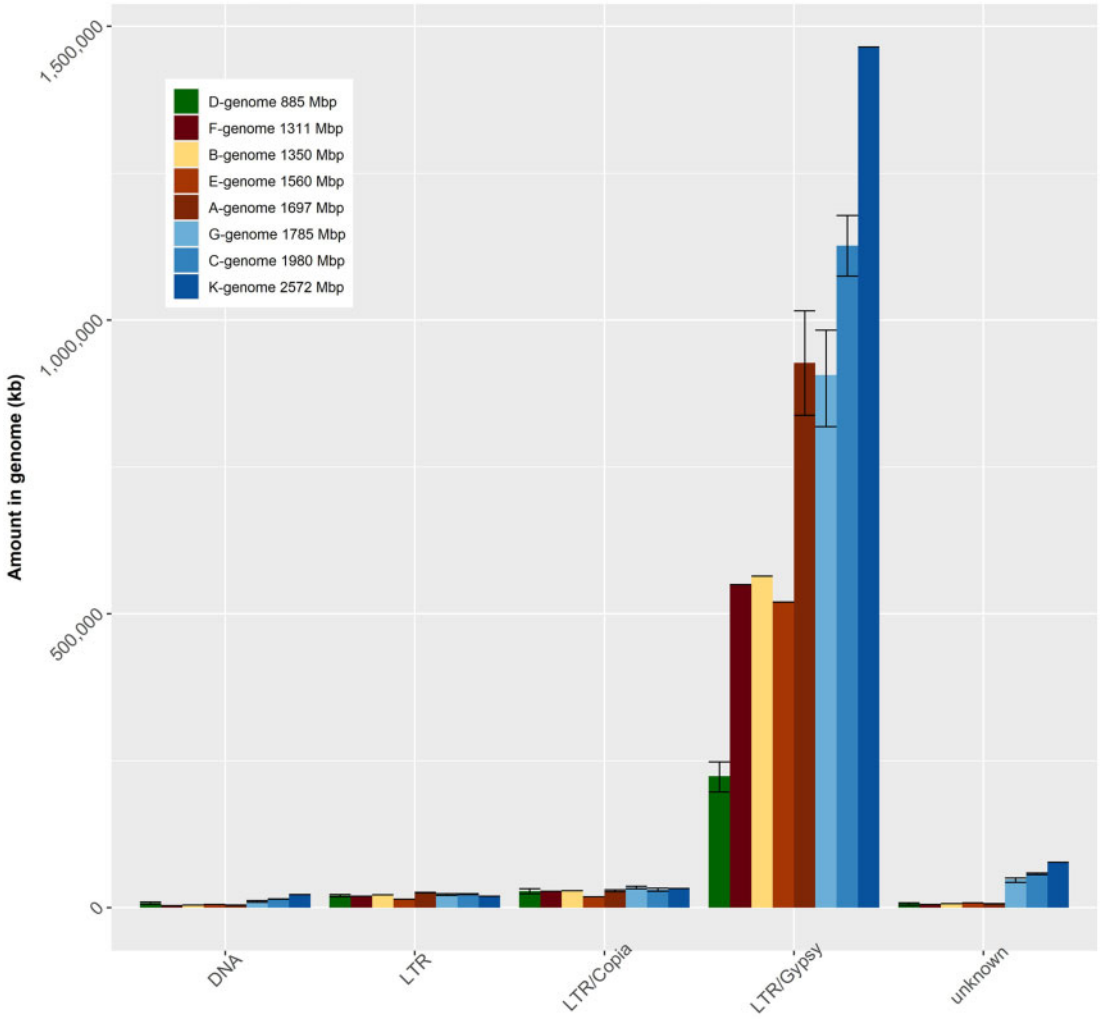

Figure 3 Average transposable element amounts in each genome group. Genome groups follow (Wang *et al.* 2018), and genome sizes for each genome group are from (Hendrix and Stewart 2005). Here, *G. anomalum* (yellow) is the sole representative of the B-genome clade.

Table 4 Reads uniquely mapped to *G. arboreum* (A2), *G. anomalum* (B1), and *G. raimondii* (D5)

Accession	Species	Total	Mapped	B1	B1 %	A2	A2 %	D5	D5 %
SRR6334584	*Gossypium davidsonii*	242,344,682	240,324,128	7,270,216	3.03	10,121,407	4.21	148,054,946	61.61
SRR8136261	*Gossypium davidsonii*	337,437,239	334,699,019	8,624,842	2.58	9,092,561	2.72	196,922,555	58.84
SRR6334602	2(A2D1)	310,722,422	295,843,181	4,904,740	1.66	159,135,266	53.79	59,248,960	20.03
SRR6334601	2(A2D3)	322,043,617	316,189,434	6,230,117	1.97	132,879,559	42.03	80,051,268	25.32
SRR3560153	*Gossypium anomalum*	305,977,284	297,638,880	250,268,307	84.08	5,231,412	1.76	718,280	0.24
SRR3560155	*Gossypium anomalum*	269,209,430	263,528,164	221,044,814	83.88	4,620,671	1.75	637,811	0.24
SRR3560156	*Gossypium anomalum*	132,066,295	128,737,188	107,684,954	83.65	2,347,574	1.82	323,087	0.25
SRR12745560	*Gossypium anomalum*	219,435,120	218,381,356	183,823,170	84.18	7,061,835	3.23	969,963	0.44
SRR8979965	*Gossypium arboreum*	478,364,261	476,168,092	5,455,280	1.15	346,423,334	72.75	1,719,046	0.36
SRR8979944	*Gossypium arboreum*	424,564,635	422,364,742	4,765,932	1.13	308,666,069	73.08	1,595,026	0.38
SRR8979925	*Gossypium arboreum*	415,343,892	414,200,416	4,494,645	1.09	302,706,188	73.08	1,290,059	0.31

B06G223900, however has no obvious sequence or structural differences, other than increased heterozygosity representing the presence of both A- and B-derived alleles. The involvement of these predicted proteins in hybrid lethality is unclear, and further research, including expression-based analyses, will be required to understand the contribution of these and/or other genes to D3-lethality in cotton.

Conclusion

The cotton genus has been the beneficiary of multiple high-quality genome sequences. While many have focused on the domesticated species, recent efforts have led to the generation of reference genomes for some of the wild representatives among the approximately 50 species in the genus (Cai *et al.* 2020; Grover *et al.* 2020, 2021; Chen *et al.* 2020; Wang *et al.* 2021). Here we report the first *de novo* sequence for a representative of the B-genome (Wang *et al.* 2018), whose members provide additional germplasm resources for both understanding and incorporating features like stress resistance and/or hybrid lethality into breeding programs. This resource will provide the foundation for future research into cotton diseases, such as blackarm (Knight 1954; Fryxell 1984), as well as provide a potential source for fiber quality improvements (Mehetre 2010) and/or fertility control among different cotton lines.

Data availability

The *G. anomalum* genome sequence and raw data are available at NCBI under PRJNA421337 and CottonGen (https://www.cottongen.org/Analysis/6607279; last accessed 9/7/21;). Supplementary files are available from figshare: https://doi.org/10.25387/g3.14787558 (last accessed 9/7/21).

Acknowledgements

The authors thank and remember the late Joshua A. Lee for his contributions to science, including improving our understanding of hybrid lethality and for creating the 2(A2D3) synthetic. We thank the Iowa State University ResearchIT unit, the BYU Fulton SuperComputer lab, the USDA-ARS, and the Mississippi State University High Performance Computing Collaboratory for computational resources and support.

Funding

The authors thank the National Science Foundation Plant Genome Research Program (Grant #1339412), the United States Dept. of Agriculture—Agriculture Research Service (Grant #58-6066-6-046), and Cotton Inc. for their financial support.

Conflicts of interest

The authors declare that there is no conflict of interest.

Literature cited

Bailly-Bechet M, Haudry A, Lerat E. 2014. "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. Mob DNA. 5: 13.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 6:11.

Beasley JO. 1940. The origin of American tetraploid *Gossypium* species. Am Nat. 74:285–286.

Bohra A, Jha UC, Adhimoolam P, Bisht D, Singh NP. 2016. Cytoplasmic male sterility (CMS) in hybrid breeding in field crops. Plant Cell Rep. 35:967–993.

Bomblies K, Weigel D. 2007. Hybrid necrosis: autoimmunity as a potential gene-flow barrier in plant species. Nat Rev Genet. 8: 382–393.

Borodovsky M, Lomsadze A. 2011. Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. Curr Protoc Bioinformatics. Chapter 4:Unit 4.6.1–10.

Cai Y, Cai X, Wang Q, Wang P, Zhang Y, *et al.* 2020. Genome sequencing of the Australian wild diploid species *Gossypium australe* highlights disease resistance and delayed gland morphogenesis. Plant Biotechnol J. 18:814–828.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, *et al.* 2009. BLAST+: architecture and applications. BMC Bioinformatics. 10:421.

Campbell MS, Holt C, Moore B, Yandell M. 2014. Genome annotation and curation using MAKER and MAKER-P. Curr Protoc Bioinformatics. 48:4.11.1–39.

Chen ZJ, Sreedasyam A, Ando A, Song Q, De Santiago LM, *et al.* 2020. Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. Nat Genet. 52: 525–533.

Cronn RC, Small RL, Haselkorn T, Wendel JF. 2002. Rapid diversification of the cotton genus (*Gossypium*: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. Am J Bot. 89: 707–725.

Du X, Huang G, He S, Yang Z, Sun G, *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.

Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, *et al.*. 2016. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Syst. 3:99–101.

Eilbeck K, Moore B, Holt C, Yandell M. 2009. Quantitative measures for the management and comparison of annotated genomes. BMC Bioinformatics. 10:67.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16:157.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20:238.

Endrizzi JE, Turcotte EL, Kohel RJ. 1985. Genetics, cytology, and evolution of *Gossypium*. In: EW Caspari, JG Scandalios, editors. Advances in Genetics. London: Academic Press, p. 271–375.

Fryxell PA. 1984. Taxonomy and germplasm resources. In: RJ Kohel, CF Lewis, editors. Cotton. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison, WI. p. 27–57.

Fryxell PA. 1992. A revised taxonomic interpretation of *Gossypium* L (Malvaceae). Rheeda. 2:108–165.

Gerstel DU. 1954. A new lethal combination in interspecific cotton hybrids. Genetics. 39:628–639.

Ghosh S, Chan C-KK. 2016. Analysis of RNA-Seq data using TopHat and cufflinks. Methods Mol Biol. 1374:339–361.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, *et al.* 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29:644–652.

Grover CE, Pan M, Yuan D, Arick MA, Hu G, *et al.* 2020. The *Gossypium longicalyx* genome as a resource for cotton breeding and evolution. G3 (Bethesda) 10:1457–1467.

Grover CE, Yuan D, Arick MA, Miller ER, Hu G, *et al.* 2021. The *Gossypium stocksii* genome as a novel resource for cotton improvement. G3 (Bethesda). doi: 10.1093/g3journal/jkab125.

Hendrix B, Stewart JM. 2005. Estimation of the nuclear DNA content of *Gossypium* species. Ann Bot. 95:789–797.

Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-genome annotation with BRAKER. Methods Mol Biol. 1962:65–95.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 12:491.

Huang G, Wu Z, Percy RG, Bai M, Li Y, *et al.* 2020. Genome sequence of *Gossypium herbaceum* and genome updates of *Gossypium arboreum* and *Gossypium hirsutum* provide insights into cotton A-genome evolution. Nat Genet. 52:516–524.

Hutchinson JB. 1932. The genetics of cotton. Part VII."Crumpled": a new dominant in Asiatic cottons produced by complementary factors. J Genet. 25:281–291.

Hutchinson JB, Silow RA, Stephens SG. 1947. The Evolution of Gossypium and the Differentiation of the Cultivated Cottons. London: Oxford University Press.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, *et al.* 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics. 30:1236–1240.

Kidwell KK, Osborn TC. 1992. Simple plant DNA isolation procedures. In: JS Beckmann, TC Osborn, editors. Plant Genomes: Methods for Genetic and Physical Mapping. Netherlands, Dordrecht: Springer, p. 1–13.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 12:357–360.

Knight RL. 1954. The genetics of blackarm resistance. J Genet. 52: 466–472.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, *et al.* 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27:722–736.

Lee JA. 1981a. A genetical scheme for isolating cotton cultivars. Crop Sci. 21:339–341.

Lee JA. 1981b. A new linkage relationship in cotton. Crop Sci. 21: 346–347.

Lee JA. 1981c. Genetics of D3 complementary lethality in *Gossypium hirsutum* and *G. barbadense*. J Hered. 72:299–300.

Li F, Fan G, Wang K, Sun F, Yuan Y, *et al.*. 2014. Genome sequence of the cultivated cotton Gossypium arboreum. Nat Genet. 46: 567–572.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25:1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al.*; 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25: 2078–2079.

Maheshwari S, Barbash DA. 2011. The genetics of hybrid incompatibilities. Annu Rev Genet. 45:331–355.

Mammadov J, Buyyarapu R, Guttikonda SK, Parliament K, Abdurakhmonov IY, *et al.* 2018. Wild relatives of maize, rice, cotton, and soybean: treasure troves for tolerance to biotic and abiotic stresses. Front Plant Sci. 9:886.

Mapleson D, Venturini L, Kaithakottil G, Swarbreck D. 2018. Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. Gigascience. 7:giy131.

Marshall DR, Thompson NJ, Nicholls GH, Patrick CM. 1974. Effects of temperature and day length on cytoplasmic male sterility in cotton (*Gossypium*). Aust J Agric Res. 25:443–447.

Mehetre SS. 2010. Wild *Gossypium anomalum*: a unique source of fibre fineness and strength. Curr Sci. 99:58–71.

Menzel MY, Brown MS. 1955. Isolating mechanisms in hybrids of *Gossypium gossypioides*. Am J Bot. 42:49–57.

Meyer VG, Meyer JR. 1965. Cytoplasmically controlled male sterility in Cotton. Crop Sci. 5:444–448.

Novák P, Neumann P, Macas J. 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics. 11:378.

Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, *et al.* 2012. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature. 492:423–427.

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, *et al.* 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 33:290–295.

Phillips LL. 1963. The cytogenetics of *Gossypium* and the origin of new world cottons. Evolution. 17:460–469.

Phillips LL. 1977. Interspecific incompatibility in *Gossypium*. IV. Temperature-conditional lethality in hybrids of *G. klotzschianum*. Am J Bot. 64:914–915.

Phillips LL, Merritt JF. 1972. Interspecific incompatibility in *Gossypium*. I. Stem histogenesis of *G. hirsutum* x *G. gossypioides*. Am J Bot. 59:203–208.

Phillips LL, Reid RK. 1975. Interspecific incompatibility in *Gossypium*. II. Light and electron microscope studies of cell necrosis and tumorigenesis in hybrids of *G. klotzschianum*. Am J Bot. 62:790–796.

Quinlan AR. 2014. BEDTools: the Swiss-army tool for genome feature analysis. Curr Protoc Bioinformatics. 47:11.12.1–11.12.34.

R Core Team 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Samora PJ, Stelly DM, Kohel RJ. 1994. Localization and mapping of the *Le1*, and *Gl2* loci of cotton (*Gossypium hirsutum* L.). J Hered. 85: 152–157.

Silow RA. 1941. The comparative genetics of *Gossypium anomalum* and the cultivated Asiatic cottons. J Genet. 42:259–358.

Smit AFA, Hubley R, Green P. 2015. RepeatMasker Open-4.0. 2013–2015.

Song L, Guo W, Zhang T. 2009. Interaction of novel Dobzhansky–Muller type genes for the induction of hybrid lethality between *Gossypium hirsutum* and *G. barbadense* cv. Coastland R4-4. Theor Appl Genet. 119:33–41.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, *et al.* 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 34:W435–W439.

Stelly DM. 1990. Localization of the *Le2* locus of cotton (*Gossypium hirsutum* L.). J Hered. 81:193–197.

Stelly DM, Lee JA, Rooney WL. 1988. Proposed schemes for mass-extraction of doubled haploids of cotton. Crop Sci. 28:885–890.

Stephens SG. 1946. The genetics of corky; the New World alleles and their possible role as an interspecific isolating mechanism. J Genet. 47:150–161.

Suzuki H, Yu J, Ness SA, O'Connell MA, Zhang J. 2013. RNA editing events in mitochondrial genes by ultra-deep sequencing methods: a comparison of cytoplasmic male sterile, fertile and restored genotypes in cotton. Mol Genet Genomics. 288:445–457.

Udall JA, Long E, Hanson C, Yuan D, Ramaraj T, *et al.* 2019. De novo genome sequence assemblies of *Gossypium raimondii* and *Gossypium turneri*. G3 (Bethesda). 9:3079–3085.

UniProt Consortium 2008. The universal protein resource (UniProt). Nucleic Acids Res. 36:D190–D195.

Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D. 2018. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. Gigascience. 7:giy093.

Vollesen K. 1987. The native species of *Gossypium* (Malvaceae) in Africa, Arabia and Pakistan. Kew Bull. 42:337–349.

Wang K, Wang Z, Li F, Ye W, Wang J, *et al.* 2012. The draft genome of a diploid cotton Gossypium raimondii. Nat Genet. 44:1098–1103.

Wang M, Li J, Wang P, Liu F, Liu Z, *et al.* 2021. Comparative genome analyses highlight transposon-mediated genome expansion and the evolutionary architecture of 3D genomic folding in cotton. Mol Biol Evol. 38:3621–3636.

Wang K, Wendel JF, Hua J. 2018. Designations for individual genomes and chromosomes in *Gossypium*. J Cotton Res. 1:3.

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, *et al.* 2017. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 35:543–548.

Weaver DB, Weaver JB Jr, 1977. Inheritance of pollen fertility restoration in cytoplasmic male-sterile upland cotton. Crop Sci. 17:497–499.

Wendel JF, Brubaker CL, Seelanan T. 2010. The origin and evolution of *Gossypium*. In: JM Stewart, DM Oosterhuis, JJ Heitholt, JR Mauney, editors. Physiology of Cotton. Netherlands, Dordrecht: Springer, p. 1–18.

Wickham H, Francois R, Henry L, Müller K. 2015. dplyr: a grammar of data manipulation. R package version 0.4.3.

Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 13:329–342.

Yu J, Jung S, Cheng C-H, Ficklin SP, Lee T, *et al.* 2014. CottonGen: a genomics, genetics and breeding database for cotton research. Nucleic Acids Res. 42:D1229–D1236.

Yuan D, Tang Z, Wang M, Gao W, Tu L, *et al.* 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.

Communicating editor: R. Wisser