Dual Domestication, Diversity, and Differential Introgression in Old World Cotton Diploids

Corrinne E. Grover [1,*], Mark A. ArickII [2], Adam Thrash [2], Joel Sharbrough [3], Guanjing Hu [4,5], Daojun Yuan [6], Samantha Snodgrass [1], Emma R. Miller [1], Thiruvarangan Ramaraj [7], Daniel G. Peterson [2], Joshua A. Udall [8], and Jonathan F. Wendel [1,*]

[1]Ecology, Evolution, and Organismal Biology Department, Iowa State University, Ames, Iowa 5001, USA
[2]Biocomputing & Biotechnology, Institute for Genomics, Mississippi State University, Mississippi, USA
[3]Biology Department, New Mexico Institute of Mining and Technology, Socorro, New Mexico 87801, 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 for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China
[6]College of Plant Science and Technology, Huazhong Agricultural University, Wuhan Hubei 430070, China
[7]School of Computing, College of Computing and Digital Media, DePaul University, Chicago, Illinois 6060, USA
[8]Crop Germplasm Research Unit, USDA/Agricultural Research Service, 2881 F&B Road, College Station, Texas 77845, USA

*Corresponding authors: E-mails: corrinne@iastate.edu (C.E.G.); jfw@iastate.edu (J.F.W.).

Accepted: 01 December 2022

Abstract

Domestication in the cotton genus is remarkable in that it has occurred independently four different times at two different ploidy levels. Relatively little is known about genome evolution and domestication in the cultivated diploid species *Gossypium herbaceum* and *Gossypium arboreum*, due to the absence of wild representatives for the latter species, their ancient domestication, and their joint history of human-mediated dispersal and interspecific gene flow. Using in-depth resequencing of a broad sampling from both species, we provide support for their independent domestication, as opposed to a progenitor–derivative relationship, showing that diversity (mean $\pi = 6 \times 10^{-3}$) within species is similar, and that divergence between species is modest ($F_{ST} = 0.413$). Individual accessions were homozygous for ancestral single-nucleotide polymorphisms at over half of variable sites, while fixed, derived sites were at modest frequencies. Notably, two chromosomes with a paucity of fixed, derived sites (i.e., chromosomes 7 and 10) were also strongly implicated as having experienced high levels of introgression. Collectively, these data demonstrate variable permeability to introgression among chromosomes, which we propose is due to divergent selection under domestication and/or the phenomenon of F_2 breakdown in interspecific crosses. Our analyses provide insight into the evolutionary forces that shape diversity and divergence in the diploid cultivated species and establish a foundation for understanding the contribution of introgression and/or strong parallel selection to the extensive morphological similarities shared between species.

Key words: cotton, domestication, introgression, *Gossypium arboreum*, *Gossypium herbaceum*.

Introduction

Domestication is an important directional and in many cases diversifying evolutionary process that transformed wild plants and animals into their modern domesticated forms. Intentional selection applied to wild populations differentiates domesticates from their progenitors on both the phenotypic and genetic levels, a process usually accompanied by an overall reduction in genetic diversity in the domesticate relative to its ancestral gene pool. In some crops, domestication has occurred independently

Significance

The cotton genus (*Gossypium*) contains four different species that were independently domesticated at least 4,000 years ago. Relatively little is understood about the diversity and evolution of the two diploid African–Asian sister species *Gossypium herbaceum* and *Gossypium arboreum*, despite their historical importance in the region and contemporary cultivation, largely in the Indian subcontinent. Here we address questions regarding the relationship between the two species, their contemporary levels of diversity, and patterns of interspecific gene flow that accompany their several-millennia history of human-mediated dispersal and contact. We validate the independent domestication of the two species and document the genomic distribution of interspecific genetic exchange.

more than once (e.g., rice and common bean; (Sang and Ge 2007; Bellucci et al. 2014; Wang et al. 2014)), resulting in convergent phenotypes with potentially divergent genetic underpinnings.

The cotton genus (*Gossypium*) provides an example of a crop having multiple, independent domestications that span both continents and ploidy levels. While the two cultivated polyploid species (i.e., *Gossypium hirsutum and Gossypium barbadense*) dominate contemporary worldwide commerce, cotton also has been domesticated twice at the diploid level. Colloquially known as the "A-genome cottons", *Gossypium arboreum* and *Gossypium herbaceum* were both domesticated during the same approximate timeframe as the polyploid species (4,000–8,000 years ago), albeit in southwestern Asia and Africa (vs. the American tropics for the polyploid species; reviewed in Wendel and Grover (2015) and Hu et al. (2021)). Although fiber quality from both A-genome cotton species is inferior to that of the tetraploids, they possess spinnable fiber and are the closest living relatives to the maternal progenitor of the polyploid species (including *G. hirsutum* and *G. barbadense*; reviewed in Wendel and Grover 2015; Hu et al. 2021).

Given their historical and modern importance as crops in parts of Africa–Asia, it is surprising that so little is known regarding their origin, domestication, and modern patterns of diversity. Although *G. herbaceum* is native to the savannahs of Southern Africa (Vollesen 1987; Wendel et al. 1989; Khadi et al. 2010), the center of early diversification was likely in Northern Africa or the Near East (Fryxell 1979). *G. herbaceum* expanded bidirectionally (east–west) through the Persian Gulf States and Indian subcontinent (Kulkarni et al. 2009; Kranthi 2018). The natural and human histories of *G. arboreum* are less clear, as no wild forms have been identified. Some have suggested that *G. arboreum* may be the derivative of an early landrace of *G. herbaceum* that became isolated due to a reciprocal translocation leading to reproductive failure in subsequent generations, that is F2 breakdown (Gulati and Turner 1929; Gerstel 1953; Hutchinson 1954; Gennur et al. 1986), although recent research indicates that the two sister species separated long prior to domestication and perhaps prior to hominin (i.e., modern and extinct human species) evolution (Wendel et al. 1989; Renny-Byfield et al. 2016; Du et al. 2018; Huang et al. 2020). While little is known about the history of *G. arboreum* prior to domestication, archeological evidence and genetic diversity analyses suggest the Indus Valley as a candidate for the origin of *G. arboreum* (Gulati and Turner 1928; Wendel et al. 2010), although this may instead represent a secondary center of diversity following initial domestication elsewhere (Hutchinson 1954; Wendel et al. 2010).

Analyses of the A-genome diploids suggest that diversity within species is low (Wendel et al. 1989; Jena et al. 2011; Page et al. 2013; Fang, Gong, et al. 2017; Du et al. 2018). Recent resequencing among predominantly Chinese accessions of *G. arboreum* (Du et al. 2018) suggests that, diversity is low among those regionally restricted domesticated accessions ($\pi = 0.0002$), approximately an order of magnitude lower than recently reported (Yuan et al. 2021) for wild accessions of the domesticated polyploid species *G. hirsutum* and *G. barbadense* ($\pi = 0.0025$ in both). This observation is similar to previous reports that diversity in the diploid species is roughly equivalent to that found in the tetraploids (Wendel et al. 1989; Stanton et al. 1994). Relative diversity between the two diploid species is unclear, with the few direct comparisons reporting conflicting results (Wendel et al. 1989; Jena et al. 2011) perhaps due to differences in germplasm evaluated and/or the markers used for diversity analysis (i.e., allozymes vs. AFLP markers, respectively).

Throughout their pre-colonial history, cultivation of the A-genome diploids has been limited to Asia (Wendel et al. 1989; Basu 1996; Guo et al. 2006; Khadi et al. 2010), and their derivatives are still grown in many Asian regions (e.g., India, Myanmar, and Thailand) (Kranthi 2018), where pests and growing conditions make these species more competitive than the polyploid cultivars. In addition, A-genome diploid cottons are also used as genetic resources for introducing stress tolerance and/or disease resistance into the commercially more important polyploid cultivars (Kulkarni et al. 2009). Finally, the A-genome diploids also are of interest in that they provide a parallel to the dual domestication of cotton at the polyploid level (Yuan et al. 2021). In an effort to clarify the species history and dual domestication of these sister taxa, we employed

high throughput DNA sequencing in conjunction with evolutionary and computational biology techniques to analyze a diverse assemblage of accessions of both species. We use this whole-genome approach to improve our understanding of the modern gene pools of these species and their interrelationships to each other.

Results

Sample Selection and Verification

We resequenced 80 *G. herbaceum* and *G. arboreum* accessions, selected to represent the diversity of the A-genome clade (supplementary Table S1, Supplementary Material online). These newly sequenced accessions averaged 38× genome equivalent coverage (18×–64×; median = 35×) of the ~1700 Mbp genomes (Hendrix and Stewart 2005), a depth suitable for accurate SNP detection and diversity analysis. In addition, we included representatives of existing resequencing datasets from both A-genome species (Page et al. 2013; Du et al. 2018; Huang et al. 2020), evaluating an additional 292 accessions whose average coverage was approximately one-third of the resequencing depth (median = 9.9×) of accessions sequenced specifically for this study. Of the 372 total accessions, 154 were excluded due to low coverage (i.e., < 10× coverage; all samples were from Du et al. 2018). Phylogenetic and principal component analysis (PCA) of the remaining 218 samples (supplementary fig. S1, Supplementary Material online) led to the exclusion of seven samples due to incorrect species assignment, suggesting sample and/or germplasm (source) misidentification, and a further four were excluded as putative hybrid and/or contaminated samples (supplementary Table S1, Supplementary Material online). Notably, the remaining samples originating from (Du et al. 2018) were distinct on both the whole-genome and genic-only PCAs (supplementary fig. S2, Supplementary Material online); these 65 were consequently excluded (supplementary Table S1, Supplementary Material online) for possible batch effects due to PCR selection (Aird et al. 2011; Jones et al. 2015; Buckley et al. 2017; Tom et al. 2017). All other samples were retained for further analyses, resulting in a dataset composed of 21 *G. herbaceum* and 99 *G. arboreum* accessions (17 and 54 newly sequenced, respectively).

Diversity and Divergence Within and Among A-genome Species

Single-nucleotide polymorphisms (SNPs) within and between A-genome species were identified using the outgroup *Gossypium longicalyx* as the reference sequence. Monomorphic, derived SNPs (relative to the ancestor, *G. longicalyx*) that were shared by all accessions of both species were excluded as uninformative. In total, 12.1 million (M) variant sites (non-ancestral) were detected and distributed evenly across the *G. longicalyx* reference (supplementary Table S2, Supplementary Material online), representing <1% of the genome. In general, individual accessions were homozygous for the ancestral (*G. longicalyx*) SNP at 50–65% of variable sites, ranging from 5.6 to 7.9 M sites per sample (supplementary Table S3, Supplementary Material online), although this may be an overestimate due to joint SNP-calling (see Methods). The number of sites fixed for the derived allele (i.e., homozygous derived) varied narrowly among samples, from 1.6 to 2.1 M sites per sample, while heterozygous sites varied more broadly, from 1.7 to 4.8 M sites per sample. While the number of homozygous reference and heterozygous sites per sample is similar between *G. herbaceum* and *G. arboreum*, the number of homozygous derived sites was generally lower for *G. arboreum* (Mann–Whitney *U*, $P = 6.157e{-}06$). Fixed differences between species are relatively rare (<3% of sites), and evenly distributed across most chromosomes. Notably, chromosomes 7 and 10 from *G. herbaceum* had an order of magnitude fewer fixed, derived sites than the other chromosomes (2,385 and 2,074 vs. 23,243–38,434 for other chromosomes; supplementary Table S4, Supplementary Material online); *G. arboreum* also shared the lack of fixed sites for chromosome 7 (619 vs. 5,824–10,693).

Site frequency spectra (SFS) were generated for both *G. herbaceum* and *G. arboreum* (fig. 1). Although the outgroup *G. longicalyx* was available to represent the ancestral state, this resulted in distortion in the unfolded SFS (supplementary fig. S3, Supplementary Material online). We therefore generated a folded SFS for both species, randomly downsampling *G. arboreum* to match the sampling rate of *G. herbaceum* (21 individuals) and replicating this for 50 iterations. Comparisons among the 50 randomly downsampled iterations of the *G. arboreum* accessions suggest general congruence in the SFS (supplementary fig. S3, Supplementary Material online); therefore, a single iteration was randomly selected for display here (fig. 1). In general, both species exhibit similar patterns in their SFS, with a lower than expected peak in rare variants, as well as an abundance of intermediate and common variants, suggesting they have experienced a severe bottleneck (e.g., domestication, but with possible contributions from habitat fragmentation/loss).

Nucleotide diversity (π) within each species was generally low and unaffected by differences in sample size (fig. 2 and supplementary Table S5, Supplementary Material online). Overall, diversity was slightly higher in the more abundantly sampled *G. arboreum* (mean $\pi = 6.7 \times 10^{-3}$, vs. 5.7×10^{-3} in *G. herbaceum*), and this observation remained true both while randomly downsampling the *G. arboreum* accessions and when comparing the distribution of differences in π among randomly selected accessions (supplementary fig. S4, Supplementary Material online). Individual chromosomes

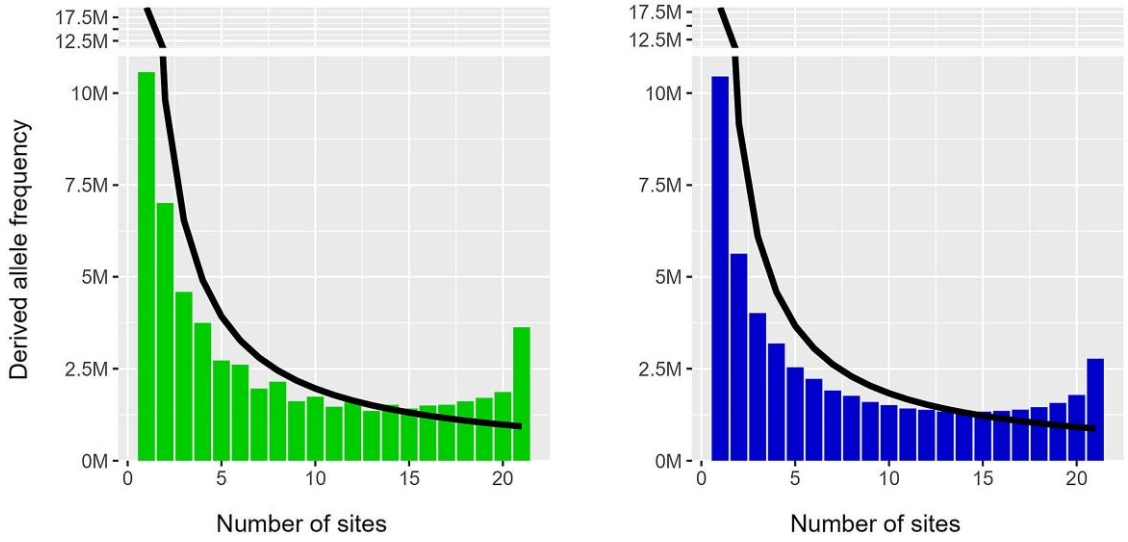


Fig. 1.—Folded site frequency distribution for *Gossypium herbaceum* (*A*; green/left) and *Gossypium arboreum* (*B*; blue/right). The black line indicates the neutral expectation based off of Watterson's theta (Hudson 2015).

also followed a general pattern of higher diversity in *G. arboreum*, although the maximum π for chromosomes F03, F04, and F06 was slightly higher in *G. herbaceum* (fig. 2). In general, diversity was chromosomally similar (supplementary fig. S4, Supplementary Material online) and correlated (supplementary fig. S5, Supplementary Material online) between *G. herbaceum* and *G. arboreum*. Watterson's theta (θ_W), however, was slightly higher in *G. herbaceum* versus nearly all downsampled iterations of *G. arboreum* (fig. 3). Because differences in sample size have been known to influence θ_W (Subramanian 2016), we computed θ_W for *G. arboreum* using the 50 randomly generated subsamples (see Methods) to determine a range in θ_W. While θ_W exhibited a range in values, it was almost always higher in *G. herbaceum* than in *G. arboreum* across all chromosomes (fig. 3) and was significantly different from a random subsample (supplementary fig. S6, Supplementary Material online). Similarly, Tajima's *D* in *G. herbaceum* was generally greater than most of the *G. arboreum* subsamples (fig. 3), although there was more overlap than exhibited by θ_W (supplementary fig. S6, Supplementary Material online). In both species, Tajima's *D* was positive (range: 0.320–0.658), likely due to the severe bottleneck experienced under domestication.

Between taxon divergence, as measured by Weir and Cockerham F_{ST}, was modest ($F_{ST} = 0.413$). Mean F_{ST} per chromosome (10 kb non-overlapping windows; fig. 4*A* and supplementary fig. S6, Supplementary Material online) varied from 0.380 on chromosome F10 to 0.490 on chromosome F05 (supplementary Table S5, Supplementary Material online), but was highly variable among chromosome windows (fig. 4*A* and supplementary fig. S6, Supplementary Material online). Nucleotide divergence between the populations was also measured by dxy (10 kb non-overlapping windows), resulting in a mean divergence (per chromosome) of 0.009–0.011 and a global mean of 0.010 (fig. 4*B* and supplementary fig. S7, Supplementary Material online; supplementary Table S5, Supplementary Material online). Using this estimate of divergence and a Malvaceae-specific mutation rate (*r*) of 4.56e–09 (De La Torre et al. 2017), we estimate divergence between these two species at approximately 1.1 Ma, much more recent than the 2.5 Ma estimate by (Renny-Byfield et al. 2016) which used a similar mutation rate (2.5 Ma using 2.6E-09). Notably, our estimates are more similar to other recent estimates (Huang et al. 2020), which report a divergence time estimate of 0.70 Ma (range = 0.40–1.40 Ma) using coalescent simulations.

Indel polymorphisms within and between species were also characterized, using the outgroup *G. longicalyx* to polarize each as either an insertion or deletion. Indels that occurred prior to species divergence, and hence were shared between *G. arboreum* and *G. herbaceum*, were discarded. Deletions generally outweighed insertions by about 50–60% within and among species (2.1 M insertions vs. 3.3 M deletions; supplementary Table S6, Supplementary Material online), although the average size (4.4 and 4.8, respectively) and size distribution of each was similar (supplementary fig. S8, Supplementary Material online). As expected, most indels (85–90%) were located in intergenic regions, and over half of genic indels (444,663 out of 867,344) were located within introns. Indels located within exons frequently resulted in frameshift mutations in gene models (75,086 indels out of 101,475; 74%), affecting just over half (20,136) of the 38,378 total genes. As with SNPs, the indel profiles of *G. herbaceum*

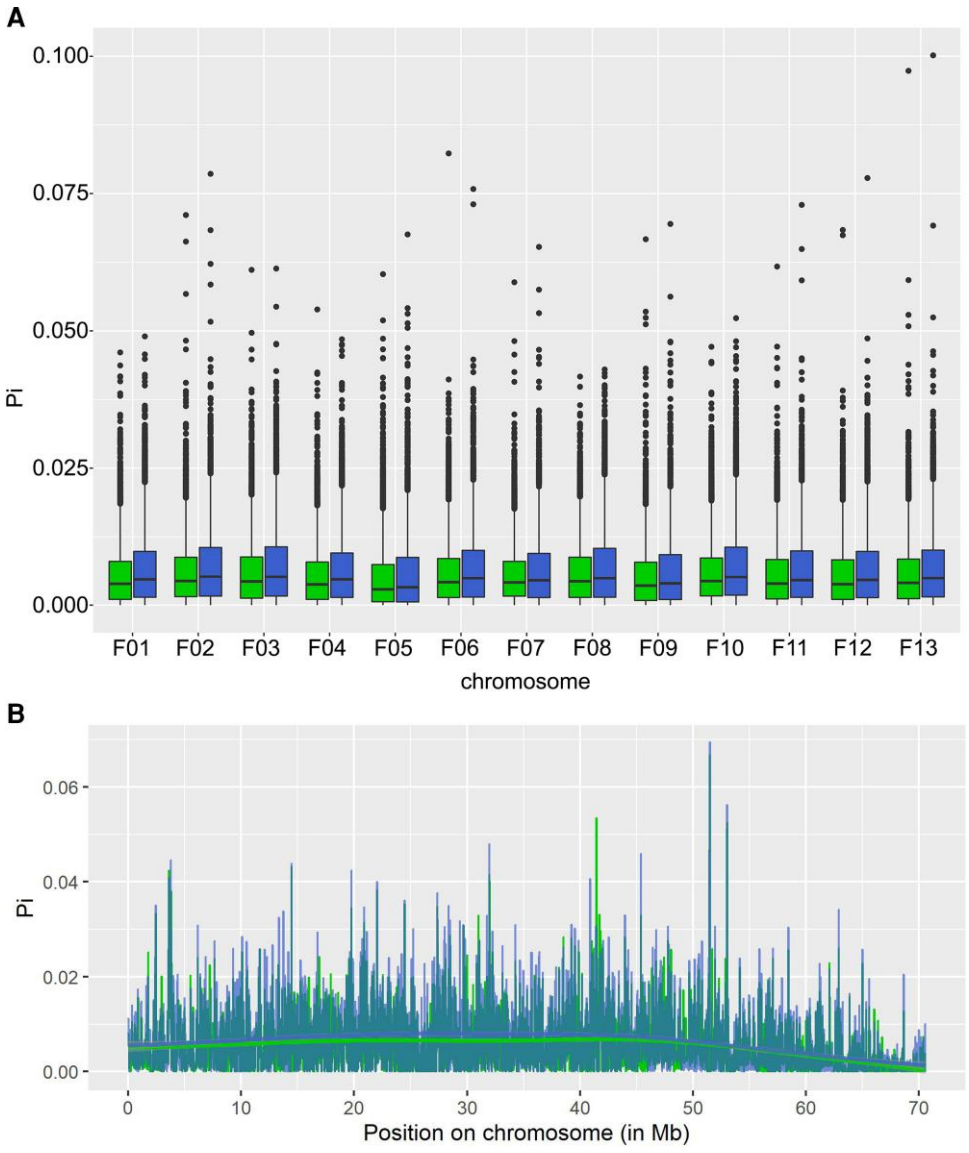


Fig. 2.—Diversity within *Gossypium herbaceum* (green) and *Gossypium arboreum* (blue). (*A*). Diversity by chromosome (10-kb windows) for *G. herbaceum* (green/left boxes) and *G. arboreum* (right/blue boxes). The chromosomal mean is depicted as a line within each box, and the lower and upper hinges correspond to the first and third quartiles, respectively. (*B*). Diversity for *G. herbaceum* and *G. arboreum* along the exemplar chromosome F09. Diversity is shown in 10-kb windows across the chromosome, and a trend line is fitted for each. In most cases, diversity is nearly identical, resulting in a darker blue-green overlap. Depictions of diversity for the remaining chromosomes can be found in supplementary figure S4, Supplementary Material online.

and *G. arboreum* were similar (supplementary Table S6, Supplementary Material online), including the difference in number of fixed indels, which was approximately three times greater in *G. herbaceum*. Notably, because the insertion and deletion rates are similar between these species (relative to the outgroup), each accession has, on average, gained ~7 Mbp of sequence and lost ~11 Mbp, leading to a net reduction in genome size due to small indels and further contributing to the divergence between species.

Phylogenetic reconstruction using genic SNPs recovered two distinct clades, one for each species (fig. 5). Phylogenetic substructure was more prominent in *G. arboreum*, which is also reflected by the mean relatedness among samples (*G. herbaceum* mean Φ=0.23; *G. arboreum* mean Φ= 0.19; see Methods). Most of the early diverging *G. herbaceum* lineages were collected on the African continent (supplementary Table S1, Supplementary Material online), with the exceptions of A1_Af (PI 630014) and A1_125 (PI 529698), which lacked collection information (both) and/or were from seed collections from Uzbekistan (A1_125). The source of the latter (A1_125) may appear to be in conflict with the African origin of *G. Herbaceum*; however,

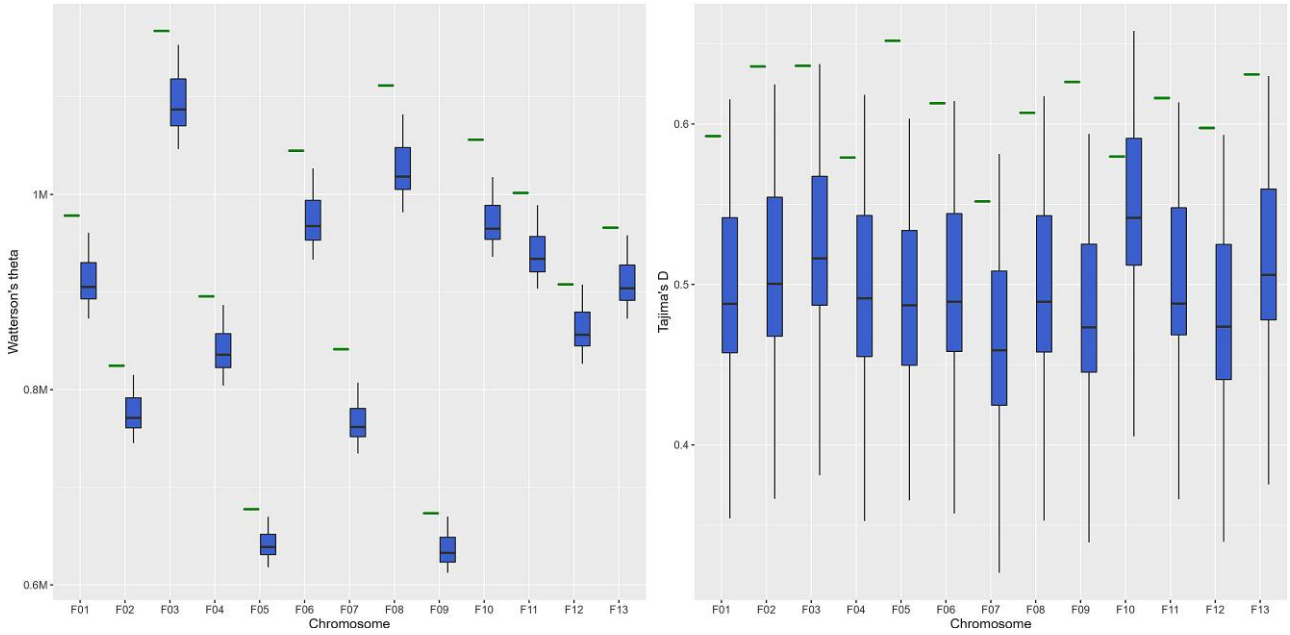


Fig. 3.—Watterson's theta (θ_W) and Tajima's *D* for *Gossypium herbaceum* (green line) and *Gossypium arboreum* (blue boxplot). Both θ_W and Tajima's *D* for *G. arboreum* were calculated using randomly subsampled accessions (50 replicates with 21 accessions each). The chromosomal mean is depicted as a line within each box, and the lower and upper hinges correspond to the first and third quartiles, respectively.

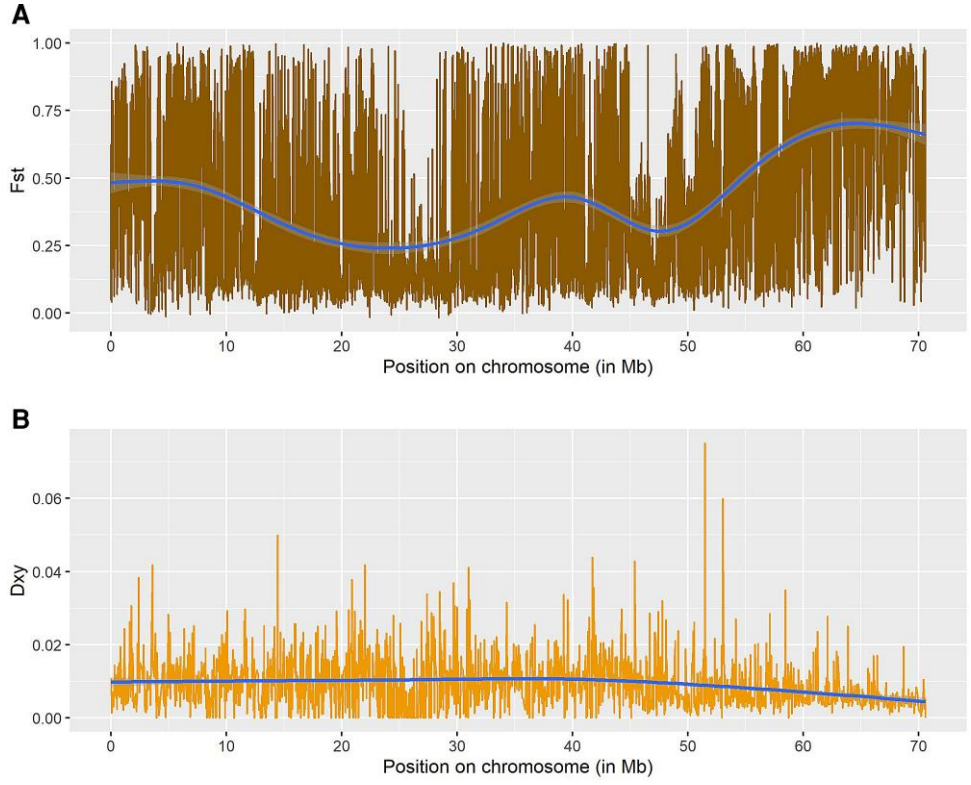


Fig. 4.—Population divergence between *Gossypium herbaceum* and *Gossypium arboreum* as measured by Weir and Cockerham F_{ST} (*A*) and interpopulation nucleotide divergence (dxy; panel B). An exemplar chromosome (F09) is depicted using 10-kb windows across the chromosome with a trend line fitted. Depictions of these divergence estimates for the remaining chromosomes can be found in supplementary figures S6 and S7, Supplementary Material online.

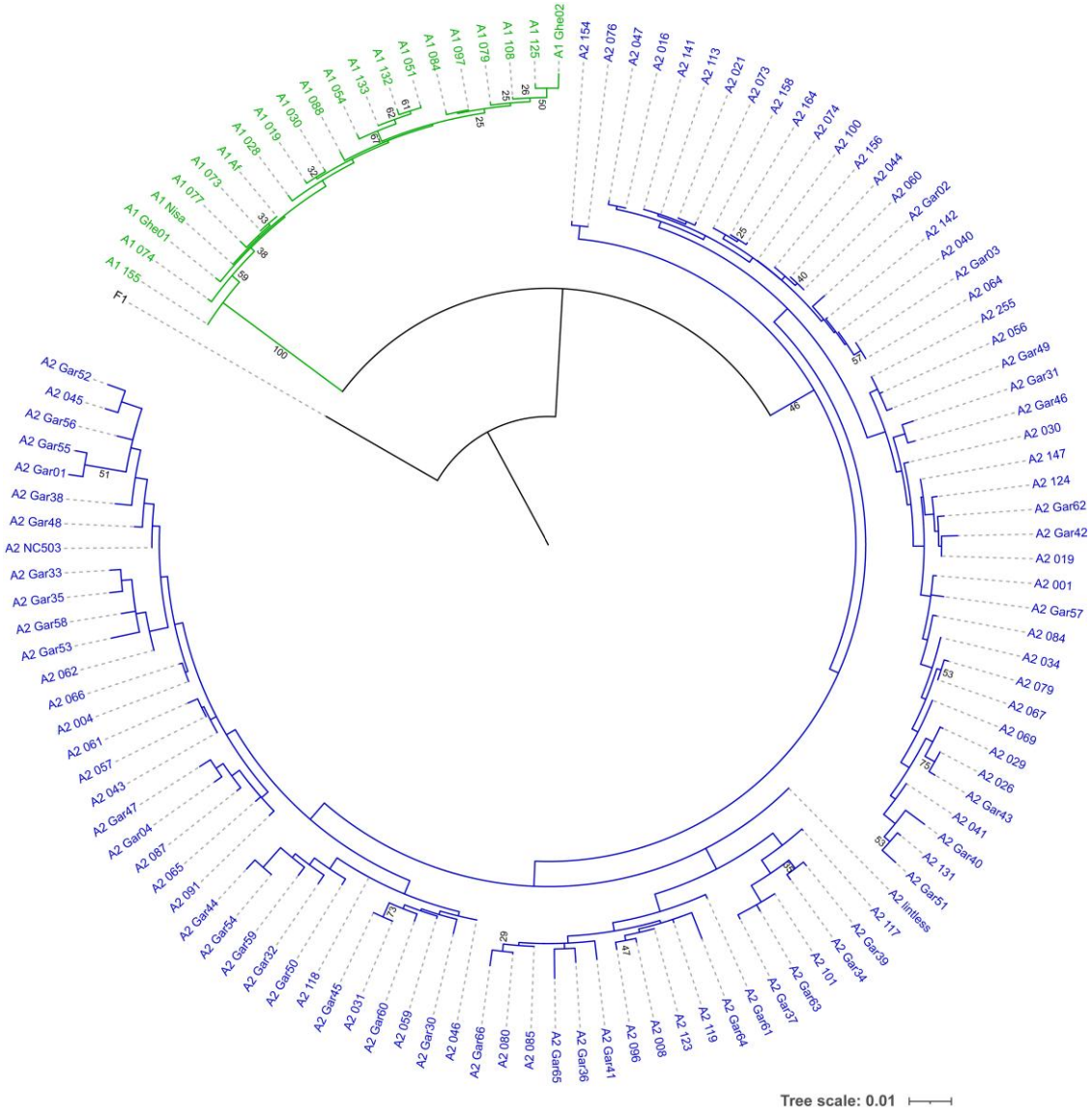


Fig. 5.—Phylogenetic analysis of accessions passing quality filters. The *Gossypium herbaceum* (A1) clade is shown in green and the *Gossypium arboreum* (A2) clade is shown in blue. Nodes with bootstrap support ≥25 are noted.

locality information in the U.S. National Plant Germplasm System (GRIN) can reflect secondary acquisition from another collection repository. Phylogenetic and geographic conflict in *G. rboretum* similarly reflect trade and secondary acquisitions in this domesticate-only species. For example, while accessions A2_073 and A2_074 have Texas, USA listed for location, the cultivar names ("Chinese naked" and "Chinese pale", respectively) indicate they may have originated in China. As expected, many of the accessions trace to the Indian subcontinent (supplementary Table S1, Supplementary Material online), which encompasses some of the major locations for cultivation of *G. arboreum*.

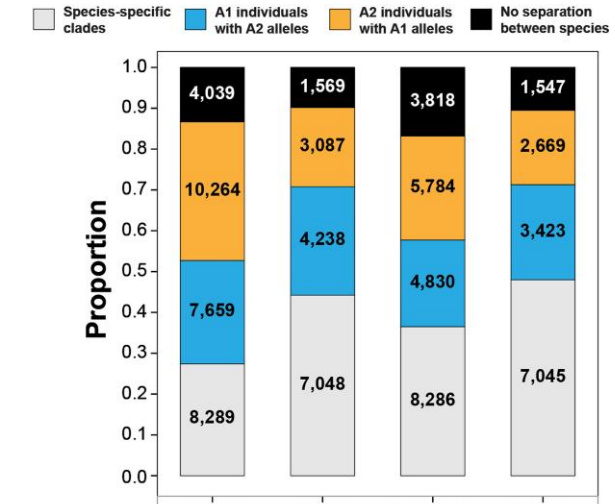
When partitioned into genomic windows composed of 50 sequential genes, phylogenetic reconstruction generally recapitulates the results of the whole genome phylogeny vis-a-vis the distinction between *G. herbaceum* and *G. arboreum*. While 478 out of the 562 windows surveyed (85%) exhibit a tree topology that is consistent with complete isolation of *G. herbaceum* and *G. arboreum,* 69 gene windows (12%; supplementary Table S7, Supplementary Material online) exhibited at least one accession nested within the alternate species, potentially signaling introgression; 15 genomic windows (3%) were excluded due to poor resolution and limited structure. Exemplar tree topologies of concordant and discordant windows are depicted in supplementary figure S9, Supplementary Material online. Most of the affected *G. arboreum* accessions contain only a single window exhibiting

possible introgression (19 out of 21 accessions); however, six of the nine *G. herbaceum* accessions (i.e., 67%) contain multiple regions exhibiting signs of introgression (median = 9.5; range = 1–37). *Gossypium herbaceum* accession A1_155 is most notable in that it is nested within *G. arboreum* in 37 regions, comprising 6.6% of the windows. While the overall phylogenetic placement for *G. herbaceum* accession A1_155 (PI 630024) is reasonable considering it is reportedly an *africanum* (hence, wild accession), the number of regions nested within its sister species may suggest it is affected by introgression. Four of the other *G. herbaceum* accessions with unusually large numbers of phylogenetically discordant windows (i.e., A1_051, A1_054, A1_132, and A1_133; median = 11.5) form a clade, suggesting that there may have been some introgression in the ancestor to these four lineages. Notably, these regions generally appear concentrated in the gene-rich distal regions of the chromosomes (supplementary fig. S10, Supplementary Material online).

These phylogenetically discordant windows were unevenly distributed, with <10% of windows affected on some chromosomes (i.e., F01, F02, F06, F12, and F13) and others with >20% of windows exhibiting discordance (i.e., F07, F08, F10). Chromosomes F12 and F13 were the only chromosomes where all phylogenetic windows exhibited a strict division between the two species (no discordance). Conversely, chromosome F07 exhibited the greatest number of discordant windows (11 of the 41 windows, or 26.8%), followed by chromosomes F08 (25.6%) and F10 (20.0% of windows). Notably, F07 and F10 also exhibit a paucity of fixed, derived sites, potentially indicating that these two are more permeable to introgression than are the other chromosomes, although we cannot disentangle the absence of gene flow from strong parallel selection.

To complement our phylogenetic analyses of 50-gene windows, we also evaluated individual gene trees for all 30,251 single-copy genes present in the dataset. We used the outgroup *G. longicalyx* to root and force bifurcation on each tree, and we subsequently compared the distribution of tree topologies that support completely independent species with those suggestive of introgression. Overall, 8,289 gene trees (27%) support the existence of distinct species-specific clades, which increased 1) to 44% when only considering trees with bootstrap support of ≥90 over at least one of the two clades, 2) to 37% when considering only trees with at least five taxa present in both clades (i.e., tree balance), or 3) to 48% when trees were required to have both ≥90 bootstrap support in at least one clade and at least five sequences in both clades (fig. 6; supplementary Table S8, Supplementary Material online).

We then tested whether genes with introgressive tree topologies were randomly distributed across the genome, as would be the case with ILS, or were clustered into



	All Trees	BS ≥ 90	≥ 5 taxa on both branches	BS ≥ 90, ≥ 5 taxa on both branches
Total trees	30,251	15,942	22,718	14,684
Total blocks	47	90	74	80
(A1,A1/A2) blocks	16	55	30	45
(A2,A1/A2) blocks	31	35	44	35

Fig. 6.—Bifurcation analysis of individual gene tree topologies. The number and proportion of gene trees for *Gossypium herbaceum* (A1) and *Gossypium arboreum* (A2) for all 30,251 single-copy genes. Four distinct tree topologies are possible: 1) species-specific clades (light gray/bottom number); 2) A1 individuals that harbor A2 alleles for a given gene (blue/second from bottom); 3) A2 individuals that harbor A1 alleles for a given gene (orange/second from top); and 4) no separation between species (black/top). Numbers and proportions of each respective topology are provided under four filtering methods: 1) No filter, 2) at least one of the clades was required to have ≥90 bootstrap support, 3) a minimum of five individuals in both clades, and 4) both bootstrap support filter and tree-balance filters combined. The table below the graph indicates, for each filtration level: 1) the total number of trees present, 2) the number of significant blocks exhibiting introgressive-like topologies, and the number of blocks suggesting 3) *G. arboreum* introgression into some accessions of *G. herbaceum* (A1, A1/A2), or 4) *G. herbaceum* introgression into some accessions of *G. arboreum* (A2, A1/A2).

physical blocks. We evaluated block size in randomly ordered genes (with replacement; see Methods) and used these bootstrap replicates to estimate the random spatial distribution of introgressive-like gene tree topologies. Although blocks as large as 15 adjacent genes could arise by chance (frequency >0.002), our analysis identified 47 uninterrupted blocks of introgressive tree topologies >15 genes (fig. 6), including one as long as 35 genes (F09, supplementary fig. S11, Supplementary Material online). These blocks were unevenly distributed between species (fig. 6), with 16 blocks characterized by possible *G. arboreum* introgression into some accessions of *G. herbaceum* (A1, A1/A2) and 31 blocks of possible *G. herbaceum* introgression into some accessions of *G. arboreum* (A2, A1/A2). Notably, however, the inferred introgression is bidirectional

and includes only a subset of accessions. In total, these blocks contain 997 genes representing 3.3% of genes. Subsequent filtering of the dataset universally led to an increase in inferred introgressive blocks, due to the removal of intervening trees that may comprise noise. For example, in our bootstrap-filtered dataset (i.e., at least one of the two subtending nodes had ≥90 bootstrap support; 15,942 trees), we identified 90 blocks of putative introgression (fig. 6), which contain 2,364 genes (14.8% of bootstrap-filtered genes) and a maximum block length of 102 genes (F05). Although these are also unevenly divided with respect to species, that is 55 blocks are consistent with *G. arboreum* introgression in *G. herbaceum* (A1, A1/A2) versus 35 consistent with the converse (A2, A1/A2), the bias in this filtered dataset suggests greater *G. arboreum* introgression into *G. herbaceum*, opposite the pattern in the unfiltered data (16 vs. 31, respectively). Interestingly, our tree balance-filtered dataset (comprising 22,718 trees; fig. 6), where at least five taxa are represented on both subtending branches, likewise identified an increased number of introgression blocks relative to the unfiltered data (i.e., 74 blocks containing 1,553, or 6.8% of genes); however, the bias in distribution between species paralleled that of the unfiltered dataset. That is, fewer blocks (30) exhibited evidence of *G. arboreum* introgression into *G. herbaceum* (A1, A1/A2) than exhibited evidence of *G. herbaceum* introgression into *G. arboreum* (A2, A1/A2; 44 blocks), similar to the unfiltered dataset. When these filters are combined (i.e., ≥90 bootstrap support and 5+taxa; 14,684 trees), the 80 introgression blocks (2,225 genes, with a maximum block length of 97 genes on F05) revert to bias toward *G. arboreum* introgression into *G. herbaceum* (45 vs. 35 blocks in the converse), perhaps due to the more stringent bootstrap filter resulting in higher tree dropout. Although the relative amount of introgression in *G. herbaceum* versus *G. arboreum* is unclear in this analysis, the presence of these numerous blocks for which trees are incongruent with the species-tree support the presence of bidirectional introgression in these species. Notably, these topological blocks are both larger than expected by chance alone and are longer than estimates of linkage disequilibrium (LD) in the A-genome (Li et al. 2018; Iqbal et al. 2021), indicating that they are unlikely to arise from ILS alone.

Synonymous Substitution Rates and Population Structure Suggest Little Interspecific Contact

Population structure analysis reveals two to three populations (fig. 7), one solely containing *G. herbaceum* accessions, and 1–2 populations comprising *G. arboreum,* depending on method (i.e., STRUCTURE/fastStructure vs. LEA, see Methods). Congruence between the two methods is high, with the major difference being the presence of substructure in the *G. arboreum* population using LEA. Congruent with the PCA, this substructure distinguishes the previously sequenced (and primarily Chinese) *G. arboreum* samples from those sequenced here. Notably, LEA also detects a higher proportion of admixture between *G. herbaceum* and *G. arboreum* accessions than either STRUCTURE or fastStructure (despite the latter including all SNPs, like LEA), which may reflect phenomena such as lineage sorting or introgression. Notably, *G. herbaceum* accession A1_155, which had the greatest number of windows indicating possible introgression, is highlighted by both STRUCTURE and LEA as containing *G. arboreum* sequence, as is *G. herbaceum* accession A1_132, albeit to a lesser degree. STRUCTURE analysis including all samples (supplementary fig. S12, Supplementary Material online) confirms the species misidentifications suggested by PCA, as well as the distinctiveness of the *G. arboreum* accessions sequenced in Du et al (2018).

Genome-wide synonymous substitution rates (d_S) between *G. herbaceum* and *G. arboreum* were estimated for 562 genome windows each containing 50 orthologous genes (supplementary Table S9, Supplementary Material online; supplementary fig. S13, Supplementary Material online) for all samples. Accessions with ≥70% ambiguity (i.e., "N") were removed from the analysis. More stringent filters were also tested and gave similar results, albeit with a lower estimated d_S (supplementary Table S9, Supplementary Material online). Notably, accessions that were considered putative hybrid and/or contaminated were easily spotted due to high or low d_S values (supplementary figs. S14 and S15, Supplementary Material online). While the low d_S values are consistent with mislabeled species and/or introgression from the sister taxon, those samples with excessive d_S values (i.e., *G. herbaceum* accessions A1_037 and A1_148 and *G. arboreum* accession A2_038) are likely introgressed with more distant species. Excluding these samples and those not passing previous quality filters (QC accessions, filtered as per methods and noted in supplementary Table S1, Supplementary Material online), the overall mean d_S between *G. herbaceum* or *G. arboreum* (supplementary Table S9, Supplementary Material online) was smaller than previously estimated from ~7,000 individual genes (d_S = 0.0088 vs. 0.0132 from (Renny-Byfield et al. 2016)). Notably, the 95% confidence interval (CI) was also broader than previously reported (Renny-Byfield et al. 2016), ranging from d_S = 0.0031–0.0198 (vs. d_S = 0.0127–0.0137), which may reflect the substantially higher sampling in the present analysis (21 *G. herbaceum* and 99 *G. arboreum* accessions, vs. two accessions each in (Renny-Byfield et al. 2016)).

Mean d_S for each chromosomal window was generally close to the genome-wide mean (i.e., within the 95% CI), although 13 windows of excess d_S (i.e., outside of the genome-wide 95% CI) were observed and a single window

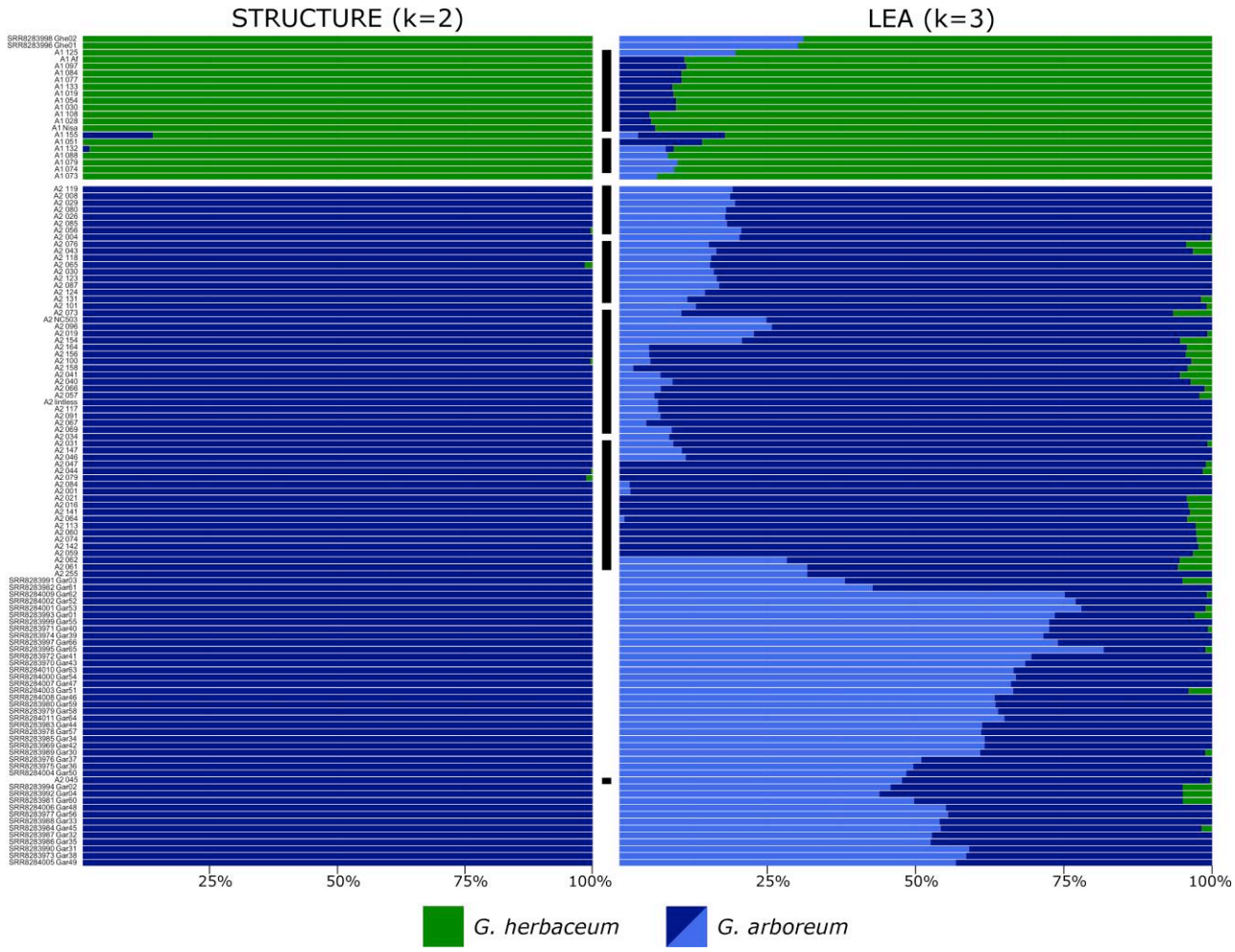


Fig. 7.—STRUCTURE (left) and LEA (right) analysis of *Gossypium herbaceum* (green) and *Gossypium arboreum* (blue). Newly sequenced accessions are noted with a black bar. Population optimization (see Methods) for STRUCTURE recovered only two populations ($K = 2$) split along species lines, whereas LEA recovered three populations ($K = 3$): one *G. herbaceum* population (green) and two *G. arboreum* populations (blue). While both STRUCTURE and LEA are based on the same underlying algorithms, LEA appears more sensitive to lineage sorting and/or introgression. A high-resolution version of this image is available at https://github.com/Wendellab/A1A2resequencing and accession details are found in supplementary Table S1, Supplementary Material online.

with reduced d_S (supplementary Table S10, Supplementary Material online). These windows are represented on approximately half of the chromosomes (6 of 13) and are frequently bordered by windows with far lower d_S. Individual genome-wide interspecific d_S estimates largely mirror the overall mean, ranging from 0.0006 to 0.0370 among windows and accessions. While all accessions appear to follow similar patterns of d_S variation across the genome (supplementary fig. S13, Supplementary Material online), ~22% of genomic windows (125 out of 562) have at least one accession whose average d_S is outside the 95% CI for the genome (supplementary Table S10, Supplementary Material online). All accessions exceed the d_S CI in at least 13 genomic windows (range: 13–44), although the median among these 125 windows is two accessions per window (range: 1–120 accessions). While d_S values below the CI were more frequently observed (67%, or 84 out of 125 windows, with 1 + accession below the genome-wide d_S), windows exceeding the d_S CI (41 windows, 33%) were generally represented by more accessions (median = 5, vs. median = 2 for low d_S windows). Notably, while 10 genomic windows exhibit significantly high d_S for >75% of accessions, only a single window has >75% of accessions with significantly low d_S (supplementary Table S10, Supplementary Material online). Furthermore, while all accessions exhibit excessively high d_S at least once, only 116 of the 120 QC accessions exhibited excessively low d_S. That is, four of the most wild accessions (i.e., A1_074, A1_073, A1_Af, and A1_Nisa; supplementary Table S10, Supplementary Material online) never exhibited excessively low d_S, perhaps indicating a lack of post-speciation interspecific contact.

Discussion

The two extant species of subgenus *Gossypium* (colloquially, the A-genome cottons) have been of great interest because they historically have been important as sources of textile fiber and because of their status as the closest relatives of the extinct A-genome donor of polyploid cottons. Disentangling the history of the A-genome cottons, including their species delimitation and intraspecific relationships, has historically been challenging due to their complex, overlapping morphologies (Fryxell 1979; Wendel et al. 1989) and putative history of introgression (Wendel et al. 1989). These challenges have led to germplasm misidentification (Wendel et al. 1989), which is evidenced here in the form of five samples misidentified as *G. herbaceum* and two misidentified as *G. arboreum* (noted in supplementary Table S1, Supplementary Material online). Genetically and cytogenetically, however, these species are distinct and exhibit evidence of reduced interspecific F_2 hybrid viability (Silow 1944; Stephens 1950; Gerstel 1953; Menzel and Brown 1954; Phillips 1961), due in part to a putatively isolating chromosomal translocation (Gennur et al. 1986). Because wild forms of *G. arboreum* are unknown and because wild forms of *G. herbaceum* are geographically disjunct from regions of cultivation (Saunders and Others 1961; Fryxell 1979; Vollesen 1987), this cytogenetic difference has historically caused some to question the independent evolution of these species (Hutchinson 1954, 1959), rather suggesting that *G. herbaceum* subspecies *africanum* represents the ancestor to both modern *G. herbaceum* and all of *G. arboreum*. These arguments have been refuted based on observations that the reconstructed divergence time between *G. herbaceum* and *G. arboreum* (Wendel et al. 1989; Page et al. 2013; Renny-Byfield et al. 2016; Huang et al. 2020) predates human agronomic innovation, typically by more than two orders of magnitude. Indeed, our estimates are similar to those previously reported, suggesting that these species diverged approximately >1 million years before present (ybp), well before domestication (circa 5,000 ybp) and comparable to previous estimates, including allozymes (1.4 Myr; (Wendel et al. 1989)), cpDNA (715,000 years; (Chen et al. 2016)) and resequencing (400,000–2.5 Myr; (Page et al. 2013; Renny-Byfield et al. 2016; Huang et al. 2020). Furthermore, phylogenetic reconstruction of both species using the outgroup *G. longicalyx* recovers a topology that clearly delineates all *G. herbaceum* accessions from *G. arboreum* and does not suggest a progenitor–derivative relationship between wild *G. herbaceum* and *G. arboreum*.

Independent evolution of *G. herbaceum* and *G. arboreum* is also supported by the prevalence of fixed homozygous derived sites in both species, 1.6–2.1 M in each, with the mean number of fixed, derived sites in *G. arboreum* slightly exceeding that in *G. herbaceum*. Fixed indels that differentiate species are likewise prevalent, with ~37,000 indels fixed in *G. herbaceum* and 6,800 different indels fixed in *G. arboreum*. While *G. arboreum* has a much lower indel fixation rate than *G. herbaceum* in the present analysis, we note that the sampling of *G. arboreum* was approximately five times greater than *G. herbaceum* and therefore the threshold to achieve fixation was greater. Supporting this is the observation that the number of differentiating indels (disregarding fixation status) is similar between species, with *G. arboreum* having slightly more indels than *G. herbaceum* (4.5 M vs. 4.4 M, respectively). Notably, nucleotide diversity was similar between the two species, that is 0.0057 and 0.0067 for *G. herbaceum* and *G. arboreum*, respectively, which also does not support a founder effect of *G. arboreum* being derived from *G. herbaceum*. When estimating diversity by omitting invariant sites, as has been previously calculated for other cotton species (fig. 2 legend), estimates of diversity (i.e., *G. herbaceum* = 0.0022 and *G. arboreum* = 0.0024) are similar to that found among wild or semi-wild and domesticated accessions of *G. barbadense* (0.0021), and greater than the diversity found within the wild-to-domesticated continuum surveyed in *G. hirsutum* (0.0017; (Yuan et al. 2021)). Interspecific divergence between the two species was modest, giving a weighted F_{ST} between *G. herbaceum* and *G. arboreum* (0.4130) similar to that between the species *G. mustelinum* and *G. ekmanianum* (0.4900; (Yuan et al. 2021)), polyploid species whose evolutionary independence is clear. These analyses represent the first direct comparison of diversity and divergence using modern techniques and diverse accessions of both species, collectively indicating recent divergence of *G. herbaceum* and *G. arboreum* followed by independent domestication.

Although we find substantial evidence consistent with independent evolution and likely domestication, we also confirm and elaborate in more detail the previous inferences of post-speciation, bidirectional, interspecific contact (i.e., introgression). While phylogenetic reconstruction based on 50-gene windows typically resulted in a clear division between species (478/562 windows, 85%), at least 12% of windows (69/562) contain topologies consistent with introgression (*i.e.*, the inclusion of one or few accessions with the alternate species). Observations of putative introgression were reiterated when considering colinear blocks of single gene trees that share topologies consistent with introgression in one or more species, most of which exceed estimates of LD for A-genome cottons. Importantly, these analyses identify non-recombined genomic blocks consistent with introgression; introgression, as opposed to ILS, is most confidently established when inferred from intact linkage blocks of introgressed chromatin. Any single discordant topology might have a number of underlying causes, but discordant chromosome blocks

provide the most compelling evidence that the discordance has arisen from secondary contact, that is interspecific gene flow in historical times. In our case, we know that the gene flow is not ancient, but more recent (Wendel et al. 1989) as the species came into contact following human-mediated dispersal from their ancestral homes under domestication (and more recently in germplasm banks). Notably, and consistent with this idea, our analyses find that *G. arboreum* alleles have only introgressed into some, but not all *G. herbaceum* lineages, whereas all *G. arboreum* lineages share a similar proportion of *G. herbaceum* alleles (supplementary fig. S11, Supplementary Material online, supplementary Table S8, Supplementary Material online). Importantly, those *G. herbaceum* lineages exhibiting little evidence of *G. arboreum* ancestry (*i.e.*, 073, 074, AfrGhe02, Nis) are the only known wild accessions sequenced here, consistent with post-domestication introgression in breeding programs or agronomic settings.

Interestingly, we observed both species-specific differences and chromosomal differences in the distribution of introgression. In general, *G. herbaceum* retains more introgressed blocks than *G. arboreum* for both the 50-gene window analysis and the dual-filtered single gene trees (median = 9.5 vs. 1 for the former and 45 vs. 35 for the latter), despite the greater sampling in the latter. Furthermore, introgression has been differentially retained among chromosomes, with some chromosomes (e.g., F13, supplementary Table S7, Supplementary Material online; or F02, supplementary fig. S11, Supplementary Material online) exhibiting no lingering evidence of introgression while other chromosomes (e.g., F07, supplementary Table S7, Supplementary Material online; or F05, supplementary fig. S11, Supplementary Material online) retain evidence of introgression in over a large portion of the chromosome. Notably, two of the three chromosomes with the highest proportion of retained introgression in our 50-gene window analysis (i.e., F07 and F10) were also exceptional in their dearth of species-specific fixed, derived sites. Studies from nearly a century ago provide potential insight into these observed differences in introgression permeability. Early research on crossing behavior in *G. herbaceum* and *G. arboreum* (Skovsted 1933; Stebbins 1945; Stephens 1949, 1950) noted F_2 breakdown in hybrids between these species consistent with underlying genetic differentiation leading to a reduction in fertility. While large structural arrangements were also known (Beasley 1942; Gerstel 1953), Stephens suggested that "small scale structural differentiations", when combined with the low crossover rate in *Gossypium*, have generally led to either gametes with near-parental structure or those which "carry deficiencies and their reciprocal duplications" (Stephens 1950). This predicts that subsequent generations would favor progeny which maximize the parental state. That is, "later generations would tend to eliminate the F_1 type and to increase the number of parental type segregates" (Stephens 1950), a consequence which Stephens notes has been generally observed with interfertile species of cotton grown in mixed cultivation. Together, these observations may highlight chromosomes and/or regions that contain factors involved in F_2 breakdown between *G. herbaceum* and *G. arboreum*, as well as indicating those chromosomes/regions that do not operate in reducing interspecific fertility and are therefore permeable to introgression. Alternatively, regions of fixed differences may indicate differential targets of selection leading either to a reduction in diversity in parts of the genome. Given the general interest in speciation genetics and islands of fertility, disentangling these two avenues would be a fruitful path for future investigation.

An interesting consequence of introgression between these species is the possibility that while *G. herbaceum* and *G. arboreum* evolved independently prior to human existence, one or both may have benefited from genetic exchange during initial domestication. Consequently, while these species evolved independently and remain distinct, further research into the identity of the alleles introgressed, and their timing and direction, may reveal a shared history that may have included introgression of some favored domestication traits. Similar observations of a single domestication spread to multiple independent lineages have been made in rice (Choi et al. 2017; Choi and Purugganan 2018), where introgression of domestication alleles from *Oryza sativa* ssp *japonica* into *O. sativa* ssp *indica* and *O. sativa* ssp. *aus* resulted in three domesticated lineages originating from a single source. Our ability to disentangle these two scenarios, that is truly independent domestication versus facilitated domestication in one species via introgression of domestication alleles, will require further research into the nature and timing of the introgressed blocks and a better understanding of domestication in the diploid cottons in general.

Cotton is an interesting model for domestication in that four species were domesticated in parallel at two different ploidy levels, providing a naturally replicated experiment for understanding convergent paths of crop evolution. Research into the evolution and domestication of the polyploid cultivars has been extensive and has yielded valuable insights in this regard (Applequist et al. 2001; Chaudhary et al. 2008; Hovav et al. 2008; Rapp et al. 2010; Said et al. 2013; Hu et al. 2014, 2019; Fang, Gong, et al. 2017; Fang, Guan, et al. 2017; Fang, Wang, et al. 2017; Chen et al. 2020; Gallagher et al. 2020; Grover, Yoo, et al. 2020; Li et al. 2021; Yuan et al. 2021). Understanding the evolution and domestication of the diploid species, however, is complicated by the lack of wild representatives for *G. arboreum*. Notwithstanding this limitation, most studies have focused on *G. arboreum*, for

which many more accessions are available and sometimes with regional biases (Du et al. 2018), or have been limited in sampling (Renny-Byfield et al. 2016) or power of the genetic markers employed (Wendel et al. 1989). The analyses presented here combined resequencing of newly acquired accessions with existing resequencing to provide a global evaluation of diversity and domestication in the A-genome species, with special consideration for evidence that supports or refutes independent evolution of these sister taxa. From these analyses, we draw the conclusion that these species evolved independently with limited interspecific contact post-speciation, albeit with lingering questions regarding the nature of their domestication(s). Subsequently, each species acquired a level of diversification and divergence that is similar to each other and to the two domesticated allopolyploids, *G. barbadense* and *G. hirsutum* (Yuan et al. 2021). While extensive morphological similarities exist between the two A-genome diploids (Wendel et al. 1989; Stanton et al. 1994), these reflect a shared history combined with a degree of phenotypic convergence and human-mediated introgression (Silow 1944; Hutchinson 1954; Wendel et al. 1989), with chromosomal and regional barriers to the latter highlighted by the uneven distribution of introgression observed here.

Methods

Germplasm Selection and Sequencing

Based on previous assessments of diversity and with the goal of capturing as much of the A-genome gene pool as possible, we selected 25 previously unsequenced accessions from *G. herbaceum* and 56 from *G. arboreum* (supplementary Table S1, Supplementary Material online). All accessions were grown in either the greenhouse or field at Brigham Young University (BYU; Provo, Utah) or the Pohl Conservatory at Iowa State University (ISU; Ames, Iowa). Young leaves were collected and high-quality DNA was extracted at BYU using the Cetyl Trimethyl Ammonium Bromide (CTAB) method (Allen et al. 2006). PCR-free libraries were constructed and sequenced using Illumina instruments (PE150) at the Beijing Genomics Institute (BGI) or the DNA Sequencing Center (DNASC) at BYU. An average coverage of 38× genome equivalents was generated for each accession.

Existing sequencing data from these two species (Page et al. 2013; Du et al. 2018; Huang et al. 2020) were downloaded (supplementary Table S1, Supplementary Material online) from the Short Read Archive (SRA) hosted by the National Center for Biotechnology Information (NCBI). In total, 19 accessions of *G. herbaceum* and 273 accessions of *G. arboreum* were downloaded, most with relatively low (<10× average genome equivalent) coverage (Du et al. 2018).

Read Mapping and SNP Inference

Raw reads were mapped to the phylogenetic outgroup *G. longicalyx* genome (Grover, Pan, et al. 2020) using bwa v0.7.17-rgxh5dw (Li and Durbin 2009) from Spack (Gamblin et al. 2015). SNPs were called using the software suite provided by Sentieon (Kendig et al. 2019) (Spack version sentieon-genomics/201808.01-opfuvzr) and following the DNAseq guidelines. This pipeline is an optimization of existing methods, such as GATK (McKenna et al. 2010), and includes read deduplication, indel realignment, haplotyping, and joint genotyping. Parameters for mapping and SNP-calling follow standard practices, and are available in detail at https://github.com/Wendellab/A1A2resequencing.

Previous results (Yuan et al. 2021) suggest that lower coverage datasets lack robustness and reproducibility. Therefore, SNP coverage for each accession was calculated by vcftools (Spack version 0.1.14-v5mvhea) (Danecek et al. 2011), and samples with insufficient depth (i.e., <10× average coverage for SNP sites present in >90% of samples) were removed from further analyses. SNP sites with more than two alternative nucleotides were excluded, and a minimum average read depth of 10, a maximum average read depth of 150, and a minor allele frequency of 5% were required for a site to be retained. We checked our filtered data for violations in Hardy–Weinberg Equilibrium (HWE), which indicates a general robustness of the filtered data, finding only ~13% of sites violate the assumption HWE with an excess of heterozygosity (Chen et al. 2017). For the purposes of PCA and phylogenetics (see below), all sites with indels or missing data were excluded. The outgroup (*G. longicalyx*) was removed from the VCF for PCA, and all sites monomorphic among the A-genome diploids were removed as uninformative. All filtering was completed in vcftools (Danecek et al. 2011), and specific parameters are available at https://github.com/Wendellab/A1A2resequencing.

SNP and Indel Analyses

Gene-associated SNPs were evaluated by intersecting the filtered VCF with the relevant feature (e.g., exon, intron, etc.) from the *G. longicalyx* annotation (Grover, Pan, et al. 2020) hosted by CottonGen (Yu et al. 2014). In each case, the Unix command grep was used to recover only the targeted feature(s), and intersectBed from bedtools2 (Spack version 2.27.1-s2mtpsu) (Quinlan 2014) was used to recover only SNP sites contained within those regions. Putative effects of each SNP (relative to the outgroup, *G. longicalyx*) were calculated by passing the entire filtered VCF to SNPEff (Cingolani, Platts, et al. 2012), which returned summary statistics as html. The SNPEff config file

and parameters are available at https://github.com/Wendellab/A1A2resequencing.

Indels were placed in a separate VCF file using vcftools (Danecek et al. 2011) with the "–keep-only-indels" flag. Samples that did not pass the SNP filtering were removed from the indel set. Because indels were mapped against the outgroup sequence, *G. longicalyx*, the reference state was considered ancestral, allowing the alternate state to be characterized specifically as an insertion or deletion; this was completed using "varType" from SnpSift (Cingolani, Patel, et al. 2012). Indel effects were characterized using SNPEff, as above.

Nucleotide diversity (π) was calculated in pixy v1.2.5.beta1 (Korunes and Samuk 2021) using 100 kb windows and run via Miniconda3 Spack version 4.3.30-qdauveb. Between population divergence (dxy) was also calculated using pixy in 10 kb, non-overlapping windows and specifying population of origin. F_{ST} between populations was similarly calculated in 10 kb, non-overlapping windows in pixy, also specifying the population of origin (supplementary Table S1, Supplementary Material online). Nucleotide diversity, dxy, and F_{ST} were only calculated for samples/sites passing the above filters. Site frequency spectra were computed in angsd v0.938 (Korneliussen et al. 2014) specifying "-doMaf 1 -doMajorMinor 1 -uniqueOnly -GL 2 -minMapQ 30 -minQ 20 -minInd 19" and *G. longicalyx* as the reference, and the folded SFS was computed using realSFS from angsd. Watterson's theta and Tajima's *D* were both calculated using thetasStat from angsd. Relatedness within each species was calculated in vcftools (Danecek et al. 2011) using "–relatedness2", a KING-based method (Manichaikul et al. 2010). All plots were generated in R using ggplot2 (Wickham 2016) and trendlines were fitted using mgcv::gam in ggplot2.

Synonymous Substitution Rates

Genome-wide synonymous substitution rates were calculated for windows of 50 genes each, with the last window along each chromosome containing slightly fewer genes. Two haplotypes for each accession were reconstructed (relative to the *G. longicalyx* reference) from the mapped reads using *bam2consensus* from BamBam v. 1.3 (Page et al. 2014) and requiring a minimum of five mapped reads. In constructing windows, we only used genes that had <70% missing data, to prevent short and/or phylogenetically uninformative genes from overly influencing divergence estimates. This resulted in 563 non-overlapping windows, with a mean of 42.23 windows per chromosome (range = 31–65 windows on chromosomes F02 and F05, respectively). The synonymous substitution rate (d_S) between *G. herbaceum* and *G. arboreum* was then estimated for each window by permuting all combinations of haplotypes from G. *herbaceum* and *G. arboreum* with both haplotypes from the outgroup *G. longicalyx*. This resulted in eight separate haplotype permutations for each *G. herbaceum–G. arboreum* accession pair per genomic window for a total of 112,832 permutations of each genomic window using all accessions. The total synonymous distance between *G. herbaceum* and *G. arboreum* (outgroup = *G. longicalyx*) was estimated for each permutation of each window by employing model 0 (single ω estimated for the unrooted three-taxon tree) from codeml inside Phylogenetic Analysis by Maximum Likelihood (PAML) v. 4.9j (Yang 2007). Synonymous substitution rates inferred by codeml were extracted from codeml output using a custom script (dSPermutations.py), and visualized using ggplot2 (Wickham 2016) in R v 4.05 (R Core Team 2020). R code and PAML parsing scripts are available at https://github.com/Wendellab/A1A2resequencing. Because nonfunctional genes can inflate estimates of d_S, we repeated the analysis using a series of filters with increasing stringency to iteratively remove genes based on the number of stop codons (i.e., no limit, < 4, and 0 for low, medium, and high stringency, respectively); all stringency filters removed genes with >70% ambiguity. Overall, these filters did not alter the conclusions drawn from these data, but their values are shown in supplementary Table S2, Supplementary Material online.

Divergence time between *G. herbaceum* and *G. arboreum* was estimated using a previously calculated rate of synonymous substitutions for the Malvaceae (4.56e–09 substitutions/year), which includes *Gossypium* (De La Torre et al. 2017). We estimated divergence between *G. herbaceum* and *G. arboreum* using the equation $T = d_S/(2r)$, where d_S is represented by the mean d_S between species (excluding outliers) and r is the Malvaceae-specific synonymous substitution rate. The range in divergence time was calculated using the 95% CI for each filter level.

Phylogenetics and PCA

For samples with a minimum 10× average read coverage per SNP, we generated a neighbor-joining tree using VCF-kit commit 25c7c03 (Cook and Andersen 2017) with default parameters. After pruning samples with incorrect or questionable identity, a new phylogeny was generated using maximum likelihood estimation in SNPhylo (Lee et al. 2014) and in RAxML-NG v1.1.0 (Kozlov et al. 2019). We also inferred phylogenetic trees for the 50-gene windows used for the d_S analyses ("low filter" only, which removes sequences with >70% ambiguity; supplementary Table S2, Supplementary Material online) in RAxML v8.2.12 (Stamatakis 2014). RAxML was run using the rapid bootstrapping algorithm (100 bootstrap replicates) assuming a GTRGAMMAIX model of molecular evolution, and *G. longicalyx* was specified as the outgroup to *G. herbaceum* and *G. arboreum*. Bifurcations with low bootstrap support (i.e., ≤60 bootstrap support) were collapsed into

polytomies using a custom Python script (collapseLow SupportBranches.py) available at https://github.com/Wendellab/A1A2resequencing. Putative introgression was evaluated by screening for tree topologies that contain highly supported clades composed entirely of *G. herbaceum* or *G. arboreum* and which also include every representative accession for that species.

PCA was initially conducted for all samples passing the filters described above using the R v4.0.2 (R Core Team 2020) package SNPRelate v 1.22.0 (Zheng et al. 2012). Subsequently, misidentified or putative hybrid samples were removed to compute an exon-only PCA, using the VCF generated above. Data were visualized using ggplot2 (Wickham 2016).

Individual Gene Tree Inference and Analysis

Individual gene trees were inferred by first generating gene alignments using bam2consensus from the BamBam suite (Page et al. 2014) and filtering the resulting alignments for accessions with >75% missing data and sites with >50% missingness. IQtree2 (Minh et al. 2020) was run for each alignment with 1000 bootstraps ("–alrt 10000 -B 1000 -bnni").

To characterize the distribution of gene tree topologies across the genome, we employed a bifurcation analysis, in which *G. longicalyx* was used to root gene trees and force bifurcation using the DendroPy Python module. We then evaluated the composition of the two subtending branches using a custom Python script (aTreeTopology.py, available at https://github.com/Wendellab/A1A2resequencing), which categorized gene trees into four possible groups: 1) gene trees that exhibited species-specific clades, 2) gene trees in which one clade featured only *G. arboreum* individuals and the other was a mixed *G. arboreum*/*G. herbaceum* clade, 3) gene trees in which one clade featured only *G. herbaceum* individuals and the other was a mixed *G. arboreum*/*G. herbaceum* clade, and 4) gene trees in which both clades exhibited mixed composition with no species separation. We evaluated the relative abundances of the different possible topologies as well as the phylogenetic placement of each individual accession in all 30,251 single-copy genes. To verify that this bifurcation analysis was demonstrative of patterns of speciation and introgression, we also employed a series of filters (both alone and in concert). These filters included 1) requiring at least one of the two subtending nodes to have ≥90 bootstrap support and 2) requiring both subtending lineages to have ≥5 individuals (i.e., tree balance).

Because introgression is expected to result in larger gene blocks of similar genealogy than incomplete lineage sorting (ILS), we compared the block size of the observed gene tree topologies to block sizes generated by 10,000 replicates of randomly subsampling gene tree topologies with replacement. Because ILS is expected to occur randomly across the genome, these randomly ordered topology blocks form our null expectation for tree topologies under a scenario of no introgression. The longest topology block identified in the randomly generated topology orders was 16 genes, which occurred once in each of two distinct replicates. We therefore used 16 genes as our threshold for identifying tree topology blocks longer than could arise by random chance (*i.e.*, ILS).

Population Structure

Population structure was predicted using two datasets, one containing all samples (except for the outgroup, *G. longicalyx*), including those considered mislabeled by PCA, and the other containing only samples passing quality/identity filters (see above). The larger dataset containing all samples was thinned with vcftools to 1 SNP per 10 kb, and then both were filtered with vcftools to remove loci with more than 10% missing data and individuals with more than 95% missing data. Due to capacity limitations in STRUCTURE, a subset of 10,000 loci were randomly selected from each of the filtered VCFs (Burgos et al. 2014) and subsequently converted to STRUCTURE format via plink v1.9 (Purcell et al. 2007). Population information was added to each of these STRUCTURE input files using a custom python script available from https://github.com/Wendellab/A1A2resequencing. A third STRUCTURE dataset was created to further evaluate population structure in *G. arboreum* by removing *G. herbaceum* accessions prior to STRUCTURE conversion. Custom scripts and detailed parameters are available at https://github.com/Wendellab/A1A2resequencing.

STRUCTURE v2.3.4 (Pritchard et al. 2000; Falush et al. 2007, 2003; Hubisz et al. 2009) was run on each VCF using the range $K = 1$ to $K = 5$. Each individual K was run 16 times per file (for *G. herbaceum* and *G. arboreum*, together) or 8 times (*G. arboreum* only). STRUCTURE results were compressed into ZIP archives and uploaded to STRUCTURE Harvester (Earl and vonHoldt 2012), which uses the Evanno method (Evanno et al. 2005; Gilbert 2016) to determine the optimal K. In addition, fastStructure v1.0 ((Raj et al. 2014)) was run on filtered data using the range $K = 1$ to $K = 10$. The optimal K was determined using the script provided by fastStructure. The best K for each set of results for both STRUCTURE and fastStructure was visualized using ggplot2 in R v4.0 to show membership proportion for individuals and to show which individuals had the most similar membership proportions.

A second evaluation of population structure was completed for *G. herbaceum* and *G. arboreum* using LEA (Frichot and François 2015), which implements a STRUCTURE-like admixture analysis in the R environment (here, in R v4.0). The original and filtered VCFs were

thinned via plink to include a subset of markers in approximate linkage equilibrium using "–indep-pairwise" to remove any pair of SNPs within a 50 SNP window (sliding 10 SNPs) with an allele count correlation (r^2) value greater than 0.1 (Liu et al. 2020). A subset containing only *G. arboreum* accessions was created by filtering missing data (i.e., keeping sites with <10% missing data and individuals with <95% missing data, as described above) and removing *G. herbaceum* accessions via vcftools. LEA was run 10 times per K ($K = 1$ to $K = 10$) for each dataset. The cross-entropy criterion was plotted against the number of inferred ancestral populations for each analysis, retaining results for the K-value with the minimum cross-entropy (i.e., the lowest point on the curve). As with STRUCTURE/fastStructure, the best K for each set of results was visualized with ggplot2 in R v4.0 to show membership proportion for individuals and to show which individuals had the most similar membership proportions.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgements

We thank Matthew Hufford for assistance with the population genetic analyses and our anonymous reviewers for thoughtful suggestions. We thank Justin Conover and Kenneth McCabe for greenhouse assistance. We thank the ResearchIT unit at Iowa State University for computational support. This work was supported by the National Science Foundation (to J.F.W. and J.A.U.) and the New Mexico Institute of Mining and Technology. We also used resources from the University of Colorado Boulder Research Computing Group, which is supported by the National Science Foundation (awards ACI-1532235 and ACI-1532236), the University of Colorado Boulder, and Colorado State University.

Data Availability Statement

The data used in this article are available from the Short Read Archive (under PRJNA539957) at https://www.ncbi.nlm.nih.gov/sra for sequencing data and from Github (https://github.com/Wendellab/A1A2resequencing) for code and analyses. Supplementary information is available at https://doi.org/10.6084/m9.figshare.21347568.

Literature Cited

R Core Team. 2020. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Aird D, et al. 2011. Analyzing and minimizing PCR amplification bias in illumina sequencing libraries. Genome Biol. 12:R18.

Allen GC, Flores-Vergara MA, Krasynanski S, Kumar S, Thompson WF. 2006. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. Nat Protoc. 1: 2320–2325.

Applequist WL, Cronn R, Wendel JF. 2001. Comparative development of fiber in wild and cultivated cotton. Evol Dev. 3:3–17.

Basu AK. 1996. Current genetic research in cotton in India. Genetica. 97:279–290.

Beasley JO. 1942. Meiotic chromosome behavior in Species, Species hybrids, haploids, and induced polyploids of *Gossypium*. Genetics. 27: 25–54. https://pubmed.ncbi.nlm.nih.gov/17247031/

Bellucci E, et al. 2014. Genomics of origin, domestication and evolution of *Phaseolus vulgaris*. In: Tuberosa R, Graner A, Frison E, editors. Genomics of plant genetic resources: volume 1. Managing, sequencing and mining genetic resources. Dordrecht, Springer, Netherlands. p. 483–507. doi: 10.1007/978-94-007-7572-5_20.

Buckley AR, et al. 2017. Pan-cancer analysis reveals technical artifacts in TCGA germline variant calls. BMC Genomics. 18:458.

Burgos NR, et al. 2014. The impact of herbicide-resistant rice technology on phenotypic diversity and population structure of United States weedy rice. Plant Physiol. 166:1208–1220.

Chaudhary B, et al. 2008. Global analysis of gene expression in cotton fibers from wild and domesticated *Gossypium barbadense*. Evol Dev. 10:567–582.

Chen Z, et al. 2016. Chloroplast DNA structural variation, phylogeny, and age of divergence among diploid cotton Species. PLoS One. 11:e0157183.

Chen ZJ, et al. 2020. Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. Nat Genet. 52:525–533.

Chen B, Cole JW, Grond-Ginsbach C. 2017. Departure from Hardy Weinberg equilibrium and genotyping error. Front Genet. 8:167.

Choi JY, et al. 2017. The rice paradox: multiple origins but single domestication in Asian rice. Mol Biol Evol. 34:969–979.

Choi JY, Purugganan MD. 2018. Multiple origin but single domestication Led to *Oryza sativa*. G3. 8:797–803.

Cingolani P, Patel VM, et al. 2012. Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. Front Genet. 3:35.

Cingolani P, Platts A, et al. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin). 6:80–92.

Cook DE, Andersen EC. 2017. VCF-kit: assorted utilities for the variant call format. Bioinformatics. 33:1581–1582.

Danecek P, et al. 2011. The variant call format and VCFtools. Bioinformatics. 27:2156–2158.

De La Torre AR, Li Z, Van de Peer Y, Ingvarsson PK. 2017. Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. Mol Biol Evol. 34:1363–1377.

Du X, 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.

Earl DA, vonHoldt BM. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the evanno method. Conserv Genet Resour. 4:359–361.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol. 14:2611–2620.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 164:1567–1587. https://pubmed.ncbi.nlm.nih.gov/12930761/.

Falush D, Stephens M, Pritchard JK. 2007. Inference of population structure using multilocus genotype data: dominant markers and null alleles. Mol Ecol Notes. 7:574–578.
Fang L, Gong H, et al. 2017. Genomic insights into divergence and dual domestication of cultivated allotetraploid cottons. Genome Biol. 18:33.
Fang L, Guan X, Zhang T. 2017. Asymmetric evolution and domestication in allotetraploid cotton (*Gossypium hirsutum* L). Crop J. 5: 159–165.
Fang L, Wang Q, et al. 2017. Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. Nat Genet. 49:1089–1098.
Frichot E, François O. 2015. LEA: an R package for landscape and ecological association studies. Methods Ecol Evol. 6:925–929.
Fryxell PA. 1979. Natural history of the cotton tribe. 1st edition. College Station, TX: Texas A&M University Press.
Gallagher JP, Grover CE, Hu G, Jareczek JJ, Wendel JF. 2020. Conservation and divergence in duplicated fiber coexpression networks accompanying domestication of the polyploid *Gossypium hirsutum* L. G3. 10:2879–2892.
Gamblin et al. 2015. The Spack package manager: bringing order to HPC software chaos. In: SC15: International Conference for High-Performance Computing, Networking, Storage and Analysis. pp. 1–12. doi: 10.1145/2807591.2807623.
Gennur MN, Habib AF, Kadapa SN, Goud JV. 1986. Cytogenetic studies in interspecific and intraspecific hybrids of *Gossypium herbaceum* L. And *Gossypium arboreum* L. Caryologia. 39:65–68.
Gerstel DU. 1953. Chromosomal translocations in interspecific hybrids of the genus *Gossypium*. Evolution. 7:234–244.
Gilbert KJ. 2016. Identifying the number of population clusters with structure: problems and solutions. Mol Ecol Resour. 16:601–603.
Grover CE, Pan M, et al. 2020. The *Gossypium* longicalyx genome as a resource for cotton breeding and evolution. G3. 10(5):1457–1467.
Grover CE, Yoo M-J, et al. 2020. Genetic analysis of the transition from wild to domesticated cotton (*Gossypium hirsutum* L). G3. 10: 731–754.
Gulati AN, Turner AJ. 1928. A note on the early history of cotton. Bombay (India): Indian Central Cotton Committee, Technological Laboratory.
Gulati AN, Turner AJ. 1929. 1—A note on the early history of cotton. J Text Inst Trans. 20:T1–T9.
Guo W-Z, Zhou B-L, Yang L-M, Wang W, Zhang T-Z. 2006. Genetic diversity of landraces in *Gossypium arboreum* L. Race sinense assessed with simple sequence repeat markers. J Integr Plant Biol. 48:1008–1017.
Hendrix B, Stewart JM. 2005. Estimation of the nuclear DNA content of *Gossypium* species. Ann Bot. 95:789–797.
Hovav R, Chaudhary B, Udall JA, Flagel L, Wendel JF. 2008. Parallel domestication, convergent evolution and duplicated gene recruitment in allopolyploid cotton. Genetics. 179:1725–1733.
Hu G, et al. 2014. Proteomics profiling of fiber development and domestication in upland cotton (*Gossypium hirsutum* L). Planta. 240:1237–1251.
Hu Y, et al. 2019. *Gossypium* barbadense and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton. Nat Genet. 51:739–748.
Hu G, et al. 2021. Evolution and diversity of the cotton genome. In: Rahman M-U- Zafar Y, Zhang T, editors. Cotton precision breeding. Cham: Springer International Publishing. p. 25–78. https://doi.org/10.1007/978-3-030-64504-5_2.
Huang G, 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.
Hubisz MJ, Falush D, Stephens M, Pritchard JK. 2009. Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour. 9:1322–1332.
Hudson R. 2015. A new proof of the expected frequency spectrum under the standard neutral model. PLoS One. 10(7):e0118087.
Hutchinson JB. 1954. New evidence on the origin of the old world cottons. Heredity (Edinb). 8:225–241.
Hutchinson JB. 1959. The application of genetics to cotton improvement. 1 edition. Cambridge, US: Cambridge University Press.
Iqbal MS, et al. 2021. Genetic factors underlying single fiber quality in A-genome donor Asian cotton (*Gossypium arboreum*). Front Genet. 12:758665.
Jena SN, et al. 2011. Analysis of genetic diversity, population structure and linkage disequilibrium in elite cotton (*Gossypium* L.) germplasm in India. Crop Pasture Sci. 62:859–875.
Jones MB, et al. 2015. Library preparation methodology can influence genomic and functional predictions in human microbiome research. Proc Natl Acad Sci U S A. 112:14024–14029.
Kendig KI, et al. 2019. Sentieon DNASeq variant calling workflow demonstrates strong computational performance and accuracy. Front Genet. 10:736.
Khadi BM, Santhy V, Yadav MS. 2010. Cotton: an Introduction. In: Cotton: biotechnological advances. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 1–14. doi: 10.1007/978-3-642-04796-1_1.
Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: analysis of next generation sequencing data. BMC Bioinform. 15:356.
Korunes KL, Samuk K. 2021. Pixy: unbiased estimation of nucleotide diversity and divergence in the presence of missing data. Mol Ecol Resour. 21:1359–1368.
Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 35:4453–4455.
Kranthi KR. 2018. Cotton production practices: snippets from global data 2017. The ICAC Recorder XXXVI:4–14.
Kulkarni VN, Khadi BM, Maralappanavar MS, Deshapande LA, Narayanan SS. 2009. The worldwide gene pools of *Gossypium* arboreum L. and G. *herbaceum* L., and their improvement. In: Paterson AH, editors. Genetics and genomics of cotton. New York, NY: Springer US. p. 69–97. https://doi.org/10.1007/978-0-387-70810-2_4.
Lee T-H, Guo H, Wang X, Kim C, Paterson AH. 2014. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics. 15:162.
Li J, et al. 2021. Cotton pan-genome retrieves the lost sequences and genes during domestication and selection. Genome Biol. 22:119.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25:1754–1760.
Li R, Erpelding JE, Stetina SR. 2018. Genome-wide association study of Gossypium arboreum resistance to reniform nematode. BMC Genet. 19:52.
Liu C-C, Shringarpure S, Lange K, Novembre J. 2020. Exploring population structure with admixture models and principal component analysis. Methods Mol Biol. 2090:67–86.
Manichaikul A, et al. 2010. Robust relationship inference in genome-wide association studies. Bioinformatics. 26:2867–2873.
McKenna A, et al. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.
Menzel MY, Brown MS. 1954. The significance of multivalent formation in three-species *Gossypium* hybrids. Genetics. 39:546–557. https://pubmed.ncbi.nlm.nih.gov/17247502/
Minh BQ, et al. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 37: 1530–1534.

Page JT, et al. 2013. Insights into the evolution of cotton diploids and polyploids from whole-genome re-sequencing. G3. 3:1809–1818.
Page JT, Liechty ZS, Huynh MD, Udall JA. 2014. Bambam: genome sequence analysis tools for biologists. BMC Res Notes. 7:829.
Phillips LL. 1961. The cytogenetics of speciation in asiatic cotton. Genetics. 46:77–83. https://pubmed.ncbi.nlm.nih.gov/17248036/.
Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics. 155: 945–959. https://pubmed.ncbi.nlm.nih.gov/10835412/.
Purcell S, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 81: 559–575.
Quinlan AR. 2014. BEDTools: the Swiss-army tool for genome feature analysis. Curr Protoc Bioinformatics. 47:11–12. https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471250953.bi1112s47.
Raj A, Stephens M, Pritchard JK. 2014. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics. 197:573–589.
Rapp RA, et al. 2010. Gene expression in developing fibres of upland cotton (*Gossypium hirsutum* L.) was massively altered by domestication. BMC Biol. 8:139.
Renny-Byfield S, et al. 2016. Independent domestication of two old world cotton Species. Genome Biol Evol. 8:1940–1947.
Said JI, Lin Z, Zhang X, Song M, Zhang J. 2013. A comprehensive meta QTL analysis for fiber quality, yield, yield related and morphological traits, drought tolerance, and disease resistance in tetraploid cotton. BMC Genomics. 14:776.
Sang T, Ge S. 2007. Genetics and phylogenetics of rice domestication. Curr Opin Genet Dev. 17:533–538.
Saunders JH, et al. 1961. The wild species of *Gossypium* and their evolutionary history. The wild species of *Gossypium* and their evolutionary history. https://www.cabdirect.org/cabdirect/abstract/19621601462.
Silow RA. 1944. The genetics of species development in the old world cottons. J Genet. 46:62–77. https://link.springer.com/content/pdf/10.1007%2FBF02986694.pdf.
Skovsted A. 1933. Cytological studies in cotton. I. The mitosis and the meiosis in diploid and triploid asiatic cotton. Ann Bot. 47:227–251. http://www.jstor.org/stable/43237398.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–1313.
Stanton MA, Stewart JM, Percival AE, Wendel JF. 1994. Morphological diversity and relationships in the A-genome cottons, *Gossypium arboreum* and *G. herbaceum*. Crop Sci. 34:519–527.
Stebbins GL. 1945. The cytological analysis of species hybrids. II. Bot Rev. 11:463–486.
Stephens SG. 1949. The cytogenetics of speciation in *Gossypium*. I. Selective elimination of the donor parent genotype in interspecific backcrosses. Genetics. 34:627–637. https://www.ncbi.nlm.nih.gov/pubmed/17247337.
Stephens SG. 1950. The internal mechanism of speciation in *Gossypium*. Bot Rev. 16:115–149.
Subramanian S. 2016. The effects of sample size on population genomic analyses–implications for the tests of neutrality. BMC Genomics. 17:123.
Tom JA, et al. 2017. Identifying and mitigating batch effects in whole genome sequencing data. BMC Bioinform. 18:351.
Vollesen K. 1987. The native species of *Gossypium* (Malvaceae) in Africa, Arabia and Pakistan. Kew Bull. 42:337–349.
Wang M, et al. 2014. The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. Nat Genet. 46:982–988.
Wendel JF, Brubaker CL, Seelanan T. 2010. The origin and evolution of *Gossypium*. In: Physiology of Cotton. pp. 1–18. doi: 10.1007/978-90-481-3195-2_1.
Wendel JF, Grover CE. 2015. Taxonomy and evolution of the cotton genus, *Gossypium*. In: Fang DD, Percy RG, editors. Cotton. Madison, WI, USA: Agronomy Monographs American Society of Agronomy, Inc., Crop Science Society of America, Inc., and Soil Science Society of America, Inc. p. 25–44. https://doi.org/10.2134/agronmonogr57.2013.0020.
Wendel JF, Olson PD, Stewart JM. 1989. Genetic diversity, introgression, and independent domestication of old world cultivated cottons. Am J Bot. 76:1795–1806.
Wickham H. 2016. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.
Yu J, et al. 2014. Cottongen: a genomics, genetics and breeding database for cotton research. Nucleic Acids Res. 42: D1229–D1236.
Yuan D, et al. 2021. Parallel and intertwining threads of domestication in allopolyploid cotton. Adv Sci. 2003634:1–17.
Zheng X, et al. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 28:3326–3328.

Associate editor: Maud Tenaillon