PNAS RESEARCH ARTICLE | PLANT BIOLOGY OPEN ACCESS

Evolutionary divergence of duplicated genomes in newly described allotetraploid cottons

Renhai Peng[a,1], Yanchao Xu[b,1], Shilin Tian[c,1], Turgay Unver[d,1], Zhen Liu[a,1], Zhongli Zhou[b,1], Xiaoyan Cai[b,1], Kunbo Wang[b], Yangyang Wei[a], Yuling Liu[a], Heng Wang[b], Guanjing Hu[b,e], Zhongren Zhang[c], Corrinne E. Grover[f], Yuqing Hou[b], Yuhong Wang[b], Pengtao Li[a], Tao Wang[a], Quanwei Lu[a], Yuanyuan Wang[h], Justin L. Conover[f], Hassan Ghazal[g], Qinglian Wang[h], Baohong Zhang[i,2], Marc Van Montagu[j,k,2], Yves Van de Peer[j,k,l,m,2], Jonathan F. Wendel[f,2], and Fang Liu[b,2]

Contributed by Marc Van Montagu; received June 1, 2022; accepted August 12, 2022; reviewed by Paul Gepts and Martina Strömvik

Allotetraploid cotton (*Gossypium*) species represents a model system for the study of plant polyploidy, molecular evolution, and domestication. Here, chromosome-scale genome sequences were obtained and assembled for two recently described wild species of tetraploid cotton, *Gossypium ekmanianum* [$(AD)_6$, *Ge*] and *Gossypium stephensii* [$(AD)_7$, *Gs*], and one early form of domesticated *Gossypium hirsutum*, race *punctatum* [$(AD)_1$, *Ghp*]. Based on phylogenomic analysis, we provide a dated whole-genome level perspective for the evolution of the tetraploid *Gossypium* clade and resolved the evolutionary relationships of *Gs*, *Ge*, and domesticated *G. hirsutum*. We describe genomic structural variation that arose during *Gossypium* evolution and describe its correlates—including phenotypic differentiation, genetic isolation, and genetic convergence—that contributed to cotton biodiversity and cotton domestication. Presence/absence variation is prominent in causing cotton genomic structural variations. A presence/absence variation-derived gene encoding a phosphopeptide-binding protein is implicated in increasing fiber length during cotton domestication. The relatively unimproved *Ghp* offers the potential for gene discovery related to adaptation to environmental challenges. Expanded gene families enoyl-CoA δ isomerase 3 and RAP2-7 may have contributed to abiotic stress tolerance, possibly by targeting plant hormone-associated biochemical pathways. Our results generate a genomic context for a better understanding of cotton evolution and for agriculture.

tetraploid cotton | polyploid dynamics | structure variations | adaptive evolution

Polyploidization is an important evolutionary process in many higher plants, contributing to speciation and adaptation (1–5). Allopolyploidy in particular has been considered a major evolutionary force due to the novel genomic possibilities resulting from hybridization and increased genetic variability. Approximately 1 to 2 million y ago (Mya), hybridization between geographically disjunct diploid A and D genome ancestors (2n = 26, AA and DD genome) and concomitant polyploidization generated allotetraploid cotton (2n = 52, AADD genome) (6, 7). This new allopolyploid clade subsequently diversified into the seven species recognized today [$(AD)_1$ to $(AD)_7$] (8, 9). Among them, *Gossypium hirsutum* [*Gh*, genome designation $(AD)_1$] and *Gossypium barbadense* [*Gb*, $(AD)_2$)], provide the majority of natural fiber for commercial production (10, 11). Five tetraploid cottons [$(AD)_1$ to $(AD)_5$], including the two domesticated species, have recently been sequenced using long-read technology, providing high-quality genome assemblies and genomic resources for uncovering the genetic basis of spinnable fiber (12–16). However, genome assemblies for the two most recently described wild tetraploid species, both of which are closely related to *Gh*, have not been reported. Genome sequences for these species may facilitate an improved understanding of evolution and fiber improvement in the dominant crop species, *G. hirsutum*. In fact, these species—that is, *Gossypium ekmanianum* [*Ge*, $(AD)_6$] from the Dominican Republic and *Gossypium stephensii* [*Gs*, $(AD)_7$] from the Wake Atoll near French Polynesia—are so similar to *Gh* that they have historically been confounded with wild accessions of *Gh* (8, 9). Additionally, there are no genome sequences for early domesticated or wild forms of either of the two domesticated species, precluding comparative genomic approaches to understanding cotton evolution and domestication. Among the great diversity of morphological forms spanning the wild-to-domesticated continuum in *Gh*, many of the least-improved forms occur in the Yucatan Peninsula of Mexico, including the truly wild race *yucatanense*, and the relatively unimproved race *punctatum* (*Ghp*) (17, 18). Assembling the genome of these tetraploid cottons will provide an important model system for understanding both the evolutionary consequences of polyploidy and of parallel domestication (19–21).

Significance

Wild relatives of domesticated plants provide a rich resource for crop improvement and a valuable comparative perspective for understanding genomic, physiological, and agricultural traits. Here, we provide high-quality reference genomes of one early domesticated form of the economically most important cotton species, *Gossypium hirsutum*, and two other wild species, to clarify evolutionary relationships and understand the genomic changes that characterize these species and their close relatives. We document abundant gene resources involved in adaptation to environmental challenges, highlighting the potential for introgression of favorable genes into domesticated cotton and for increasing resilience to climate variability. Our study complements other recent genomic analyses in the cotton genus and provides a valuable foundation for breeding improved cotton varieties.

Reviewers: P.G., University of California, Davis; and M.S., McGill University.

The authors declare no competing interest.

[1]R.P., Y.X., S.T., T.U., Z.L., Z. Zhou, and X.C. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: zhangb@ecu.edu, marc.vanmontagu@vib-ugent.be, yves.vandepeer@psb.vib-ugent.be, jfw@iastate.edu, or liufcri@163.com.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2208496119/-/DCSupplemental.

Published September 19, 2022.

Here, we report high-quality genome assemblies of the three tetraploid genomes, *Ge, Gs,* and *Ghp*. Using comparative genomics and phylogenomics, we reveal extensive genomic structural variations (SVs) in tetraploid cottons, and reevaluate phylogenetic relationships and divergence times within the polyploid clade. Extensive SVs are associated with phenotypic diversity, including the economically important trait, fiber length. We characterize wild gene resources that have the potential to facilitate adaptation to various abiotic and biotic stresses in domesticated cotton. These results deepen our understanding of genome evolution in polyploids and provide insight into the genetic and morphological diversity of tetraploid cottons.

Results

Assemblies and Characterization of Three Allotetraploid Cotton Reference Genomes. Three allotetraploid cotton genomes (*Ge*, *Gs*, and *Ghp*) were sequenced and assembled using a combination of sequencing technologies, including single-molecule real-time (PacBio), paired-end Illumina sequencing, and chromatin conformation capture (Hi-C). An initial assembly was generated via FALCON (22) using at least 20.83 million PacBio long reads for each cotton species (*SI Appendix*, Tables S1 and S2), and subsequently corrected using Illumina paired-end data (average 120-fold coverage). These megabase assemblies (N50 of 1.57, 1.23, and 11.49 Mb for *Ge*, *Gs*, and *Ghp*, respectively) (*SI Appendix*, Table S3) were combined with Hi-C interaction information to produce chromosome-scale scaffolds (*SI Appendix*, Fig. S1 and Tables S4 and S5), yielding final assemblies of 2.34, 2.29, and 2.29 Gb for *Ge*, *Gs*, and *Ghp*, respectively. These high-quality assemblies had scaffold N50 values of more than 107 Mb (Table 1), with 99% of bases anchored onto chromosomes and with 99% of mapped Illumina reads covering about 97% of the genomes (*SI Appendix*, Table S6). Nearly all of the 1,614 Embryophyta benchmarking universal single-copy orthologs (BUSCOs, embryophyta_odb10) (23) were complete in the *Ge* (99.2%), *Gs* (97.4%), and *Ghp* (99.3%) assemblies (*SI Appendix*, Table S7), and the long terminal repeat (LTR) assembly index (LAI score 13.7 in *Ge*, 12.8 in *Gs*, and 12.7 in *Ghp*) further indicated that these three assemblies could be considered "reference quality" (24) (*SI Appendix*, Table S8).

A total of 1,575 Mb (65%), 1,489 Mb (63%), and 1,488 Mb (65%) of sequence corresponding to transposable elements (TEs) were predicted in *Ge*, *Gs*, and *Ghp*, respectively (Table 1 and *SI Appendix*, Table S9). We identified 74,178, 74,970, and 74,520 protein-coding gene models (PCGs), respectively, of which an average of 97% had matched functional identifiers (*SI Appendix*, Tables S10 and S11). The majority (95 to 97%) of PCGs predicted in *Ge*, *Gs*, and *Ghp* had an identifiable homolog (>80% protein identity) in the published tetraploid cotton genomes (*SI Appendix*, Table S12) (15): that is, *Gh*, *Gb*, *Gossypium tomentosum* [*Gt*, $(AD)_3$], *Gossypium mustelinum* [*Gm*, $(AD)_4$], and *Gossypium darwinii* [*Gd*, $(AD)_5$]. An assessment of TE and PCG density in 1,000 equal windows per chromosome suggest a strong bias for *Copia* and PCG accumulation within 20% of the windows nearest the chromosome telomeres, having an average of 0.85-fold ($P < 10^{-16}$, Wilcox test) and 2.34-fold ($P < 10^{-16}$, Wilcox test) increase, respectively, compared to other chromosomal regions (Fig. 1 and *SI Appendix*, Fig. S2). In contrast, *Gypsy* element density exhibited an average decrease of 0.74-fold ($P < 10^{-16}$, Wilcox test) in telomeric versus other regions.

Phylogenetic Analysis of Tetraploid *Gossypium*. A maximum-likelihood phylogenetic tree was generated using 3,281 single-copy coding genes for the eight tetraploid cottons (seven species), eight representative diploid cottons [*G. herbaceum* A_1 (14), *Gossypium arboreum* A_2 (25), *Gossypium longicalyx* F_1 (26), *Gossypium australe* G_2 (27), *Gossypium thurberi* D_1 (28), *Gossypium raimondii* D_5 (10), and *Gossypium turneri* D_{10} (29)], and the phylogenetic outgroup species *Gossypioides kirkii* (*Gki*) (30). Divergence times were estimated using *Ks* values for orthologous genes (Figs. 2 *A* and *B*). The phylogeny of allopolyploid cotton is reiterated in both the A and D genome clades, as expected given their formation from A and D genome diploid antecedents. The initial divergence time within the allopolyploids was estimated at 1.80 Mya (95% CI: 1.10 to 2.72 Mya) (Fig. 2*B*), consistent with previous reports (15, 31). Except for *Gm* and *Gt*, which comprise one of the two deepest branches, tetraploid species fall into two clearly distinguished clades, each of which includes one of the two economically important domesticated cottons (upland cotton *Gh* and sea island cotton *Gb*). These two groups are hereafter referred to as the *Gh*-like and *Gb*-like clades, respectively (Fig. 2*C*), and are inferred to have diverged from each other ~0.79 Mya (95% CI, 0.49 to 1.49 Mya). *Ghp* is among the most primitive among the many forms of semidomesticated and feral derivatives found within the diversity generated by the 4,000+-y

Table 1. Features of three tetraploid cotton assemblies

Genomic features	*Ge*	*Gs*	*Ghp*
Assembly			
Genome size (Mb)	2,341.87	2,291.84	2,292.48
Scaffold number	160	243	277
Scaffold N50 (Mb)	108.06	108.2	106.96
Contig size (Mb)	2,341.51	2,291.47	2,292.40
Contig number	3,781	3,927	1,111
Contig N50 (Mb)	1.57	1.23	11.49
Gap number	3,621	3,684	834
Gap length (Mb)	0.36	0.37	0.08
Pseudochromosomes size (Mb)	2,337.03	2,272.89	2,283.07
Annotation			
TE percentage	64.86	63.01	64.89
Gene number	74,178	74,970	74,520
Genes in pseudochromosomes	74,038	73,324	74,283
Complete BUSCOs (%)	95.50	97.10	95.40

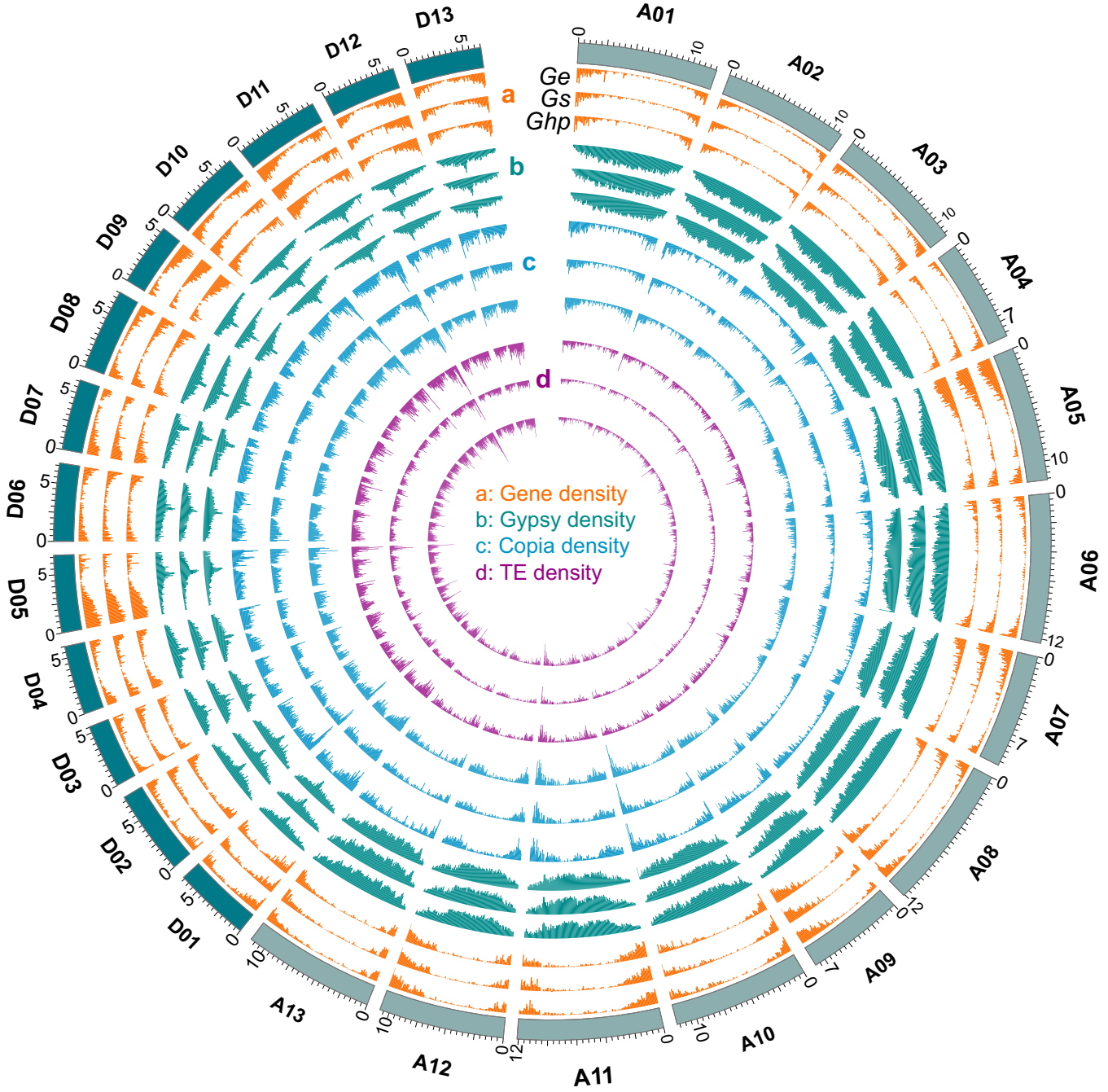

Fig. 1. Genomic feature distribution in the three cotton genomes. The scale unit of each chromosome is 10 M. Track a: Gene density indicated by gene length per megabase. Track b: Gypsy density showed the distribution of Gypsy TE length per megabase across each chromosome. Track c: Copia density showed the distribution of Copia length per megabase across each chromosome. Track d: TE density showed the distribution of DNA transposon length per megabase across each chromosome.

history of *Gh* domestication (32, 33). We observed similar divergence times between *Gb-Gd* (0.63 Mya; 95% CI: 0.37 to 1.26 Mya) and *Ghp-Gh* (0.68 Mya; 95% CI: 0.41 to 1.14 Mya), confirming earlier data indicating that the Galapagos Island endemic *Gd*, previously considered to be conspecific with *Gb*, diverged relatively recently from its mainland relatives (34). The genome sequences for the two species *Ge* and *Gs* completes the sampling of wild tetraploid *Gossypium* and, as expected from prior analyses (19–21), they fall within the *Gh*-like clade, having all diverged from their most recent common ancestor around 0.75 Mya (95% CI, 0.42 to 1.33 Mya) (*SI Appendix*, Fig. S3). With respect to diploid divergence, two branches are distinguished (Fig. 2*A*): the New World clade (D genome) and the African–Australian–Asian clade (A, G, and F genomes) (6, 7). We observed phylogenetic inconsistencies in the order of divergence within the *Gh*-like clade for the two subgenomes [Dt clade: (*Ge*, *Gs*), (*Gh*, *Ghp*)] vs. At clade: {*Ge*, [*Gs*, (*Gh*, *Ghp*)]} (Fig. 2*A*), which is reasonable given the rapid divergence exhibited by these species (8, 9).

Our results support previous inferences regarding the monophyletic origin of allopolyploid cottons (35). Although the A_T genome of tetraploid cotton is more divergent from A_2 (~1.31 Mya) than A_1 (~1.23 Mya), the range in estimates for these divergence times overlap (*SI Appendix*, Fig. S3). As

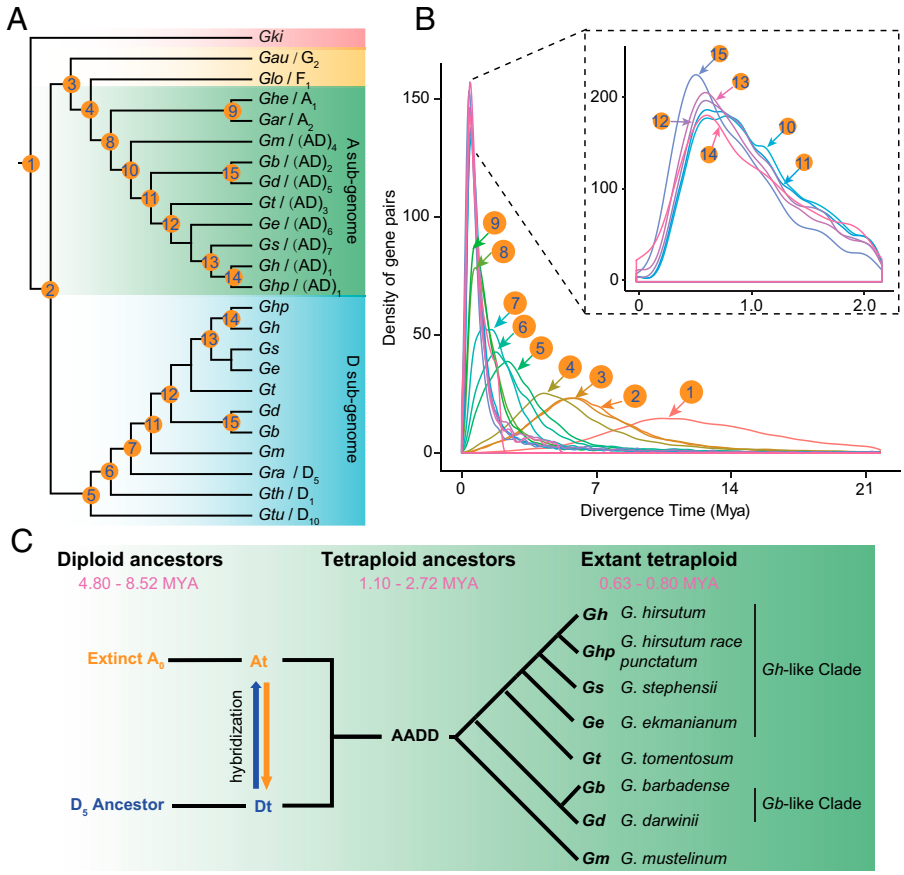

Fig. 2. Phylogenetic analysis of the *Gossypium* genomes. (*A*) Maximum-likelihood tree inferred using *G. kirkii* (*Gki*) as the outgroup. (*B*) Distribution of *Ks* values for orthologous genes among *Gossypium* genomes. (*C*) Evolution of the allopolyploid cotton clade, formed following hybridization between an extinct A_0 and ancestor of D_5.

expected, based on prior studies (7), *Gm* is the sole survivor of the earliest split in the allopolyploid species, and thus it can be used as an outgroup to evaluate subsequent evolutionary differences of the remaining allopolyploids in the *Gh*- and *Gb*-like clades. In general, the synonymous substitution rate (*Ks*) was higher for Dt homeologs than for At homeologs (*SI Appendix*, Fig. S4), consistent with a previous report (15) and possibly reflecting subgenome-specific evolutionary processes, including differences in recombination rates and selective sweeps.

Genomic SVs Occurred During the Evolution of Tetraploid *Gossypium*. SVs occur frequently during plant evolution and domestication, providing a major genetic source of phenotypic diversity (36, 37). We focused on identifying all genomic SVs ≥50 bp in length because these are the least well-characterized genetic variations and are likely to affect gene function (38, 39). By mapping the seven tetraploid *Gossypium* assembled genomes and their sequencing reads to the reference genome of *Gm* [$(AD)_4$_JGI] (15), four methods [smartie-SV (39), SVMU (40), SyRI (41), and Breakdancer (42)] were combined to identify SVs (Fig. 3) and to polarize their directionality (i.e., insertion vs. deletion relative to *Gm*). SVs were only considered when they were consistently identified by at least two methods, resulting in an average of 72,965 insertions (range 67,885 to 77,756), 63,126 deletions (range 59,663 to 65,670), and 339 inversions (range 297 to 410) (*SI Appendix*, Table S13). Presence/absence of variation (PAVs) occurred relatively frequently during cotton evolution. The lowest number of PAVs was observed in *Gt* (*SI Appendix*, Fig. S5). Notably, the domesticated polyploids (*Gb* and *Gh*) had the longest average length of PAVs among the seven tetraploid cotton genomes surveyed (*SI Appendix*, Fig. S5), yet the other five cotton accessions exhibit more PAVs with size ≥1 kb (*SI Appendix*, Table S14). Relative to the large number of species/accession-specific PAVs (range 36,476 to 75,125), fewer shared PAVs (8,277, average ratio of 6.05%) among species/accessions were observed, suggesting potential for impact by PAVs on species/accession-specific traits (relative to *Gm*) (*SI Appendix*, Fig. S6).

The number of PAVs in the Dt genome (range 61,132 to 67,223) is slightly smaller than in the At genome (range 64,875 to 78,695) for all polyploid cotton accessions except *Gh*, suggesting a higher density of PAVs in the twofold smaller Dt subgenome (*SI Appendix*, Table S13). The majority of PAVs were located in intergenic regions (70.53 to 76.81%), and fewer were in coding regions than in introns (*SI Appendix*, Fig. S7 and Table S15). PAVs overlapping exons resulted in a predicted 20,343 frameshift and 9,771 stop-codon gain or loss mutations within 11,557 predicted protein sequences (15.68% of the total) (Dataset S1), including 1,168 proteins affected in at least 4 of the genomes. The three chromosomes affected most by PAVs within genes were At05 (535 genes), At11 (428), and Dt11 (312) (*SI Appendix*, Fig. S8). Within the 8,277 PAVs shared by all accessions (relative to *Gm*), 646 protein sequences were affected by these PAVs. This branch (i.e., the internode between *Gm* and the radiation of the remaining genomes) exhibited the largest number of genic PAV-induced changes across the polyploid phylogeny (0.078) by a factor of

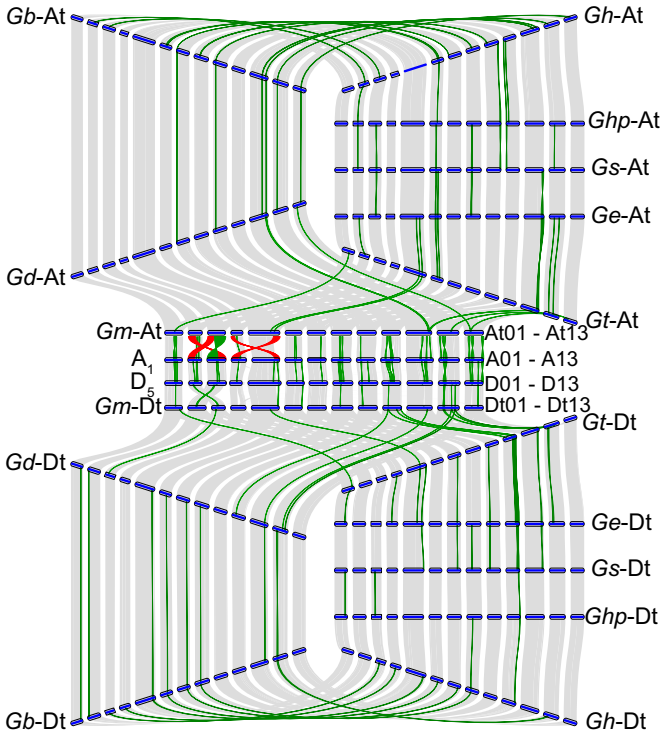

Fig. 3. Characterization of genomic variations in *Gossypium* At genome and Dt genome. Genic synteny blocks are connected by gray lines. Translocation blocks and inversion blocks are connected by red and green lines, respectively.

approximately three; on the terminal branches, species/accession-specific rates of PAV-induced genic changes ranged from 0.006 to 0.027 (*SI Appendix*, Fig. S6).

Of the SVs affecting genes, we found a 450-bp SV in a synteny block on At10 among the eight tetraploid genomes (coordinates in domesticated *Gh*; At10: 84,877,673 to 84,878,123) that resulted in a shorter version of *Ghi_A10G09231* in the *Gb*-like clade and *Gt* (Fig. 4*A*). This shorter version is phylogenetically homoplasious, possibly resulting from either a repeated truncation or a single event in a polymorphic ancestor followed by lineage sorting. This gene encodes a phosphopeptide-binding protein that is involved in fiber length, and which exhibits significantly reduced expression in fiber from cotton species missing this gene (*SI Appendix*, Fig. S9) (43). This deletion (relative to *Gm*) was present in *Gb* and *Gd*, suggesting that the deletion occurred subsequent to divergence from *Gm*. We found a large-scale inversion event (~4.48 Mb) in a synteny block on Dt04 that distinguishes the *Gh*-like clade from the *Gt* and the *Gb*-like clade (Fig. 4*B*), and thus is phylogenetically diagnosed as having occurred in the ancestor of the *Ge* through *Gh* clade. This inversion was further confirmed by mapping Hi-C data of four accessions (*Gb*, *Ge*, *Gs*, and *Ghp*) to TM-1_WHU (*SI Appendix*, Fig. S10). Notably, two genes at this inversion boundary are involved in various abiotic stress responses, in which the *Ghi_D04G05266* gene encodes calcium-dependent protein kinase and the *Ghi_D04G0499* gene encodes alcohol dehydrogenase class-P (Dataset S2) (44–48). Another large inversion (>986 kb) in a synteny block on Dt01 was present in both domesticated *Gb* and *Gh*, but not their close wild relatives. At its boundary, the gene *Ghi_D01G09866* was the homologous *BXL* in *Arabidopsis thaliana*, which encodes a β-xylosidase associated with secondary cell wall metabolism (49), suggesting a possible relationship to fiber quality; the other gene, *Ghi_D01G10141*, encodes an ethylene-responsive transcription factor, which interacts with *TPL* to regulate seed germination (Dataset S2) (50). Collectively, these SVs occurring in the evolution of tetraploid *Gossypium* led to the important phenotypic differentiation, although should be experimentally validated in the future.

Of the tetraploid cotton pan-genome made by combining our newly sequenced genomes with sequences from the five previously published tetraploid cotton species (*Gh*, *Gb*, *Gt*, *Gm*, and *Gd*) (15), a total of 96,537 gene families was detected (Fig. 5*A*). Core genes were enriched for gene ontology (GO) terms related to "regulation of biosynthetic process" and "metabolic process" (*SI Appendix*, Fig. S11), similar to previous findings from other plants (51–53). However, the number of pan-gene sets increased as additional genomes were added and did not approach a plateau (Fig. 5*A*), as for some within and between species comparisons, such as soybean and its wild relatives (51). This observation likely reflects the levels of divergence and perhaps genome size variation that exists among *Gossypium* species.

Among the *Gossypium* gene families, most (average of 59.78% or 27,484 families) were considered core families that account for an average of 68.11% of the genes, followed by approximately one-quarter of the genes that were considered "dispensable" (an average of 18,809 genes in each genome) (Fig. 5*B* and *SI Appendix*, Table S16). We also observed 6.81% of genes, on average, were species-specific, but the number and the proportion of *Gh*-like are almost twice as big as four other *Gossypium* (*SI Appendix*, Table S16). Additionally, upland and sea island cottons have slightly fewer specific genes than their closest wild relatives (i.e., *Gh* vs. *Ghp* and *Gb* vs. *Gd*) (*SI Appendix*, Table S16).

Nucleotide-binding site leucine-rich repeat (NLR) gene families mediate plant resistance to biotic stressors (54, 55). We identified a total of 3,462 to 4,312 NLR genes in each of the eight tetraploid cotton genomes (*SI Appendix*, Table S17). NLR genes were scattered across almost all chromosomes, with significantly higher numbers in Dt subgenomes than in At subgenomes ($P = 0.012$), congruent with earlier reports for five allopolyploid cotton species (15) (*SI Appendix*, Fig. S12). We observed some dense clusters appearing in A04, A11, and D11 of *Ghp* and *Ge* (*SI Appendix*, Fig. S13). Oligonucleotide probes designed for *Ge* A04 and A11 R gene clusters (*Methods*) confirmed the presence of these clusters in other tetraploid genomes (*SI Appendix*, Fig. S14). Notably, the domesticated *Gossypium* genome contained a narrower range of NLR gene domain architectures than were observed in the other genomes (*SI Appendix*, Table S17); this was especially evident in the *Gh*-like clade, where there was a trend of gradual decreasing from wild to domesticated *Gossypium*. This suggest that perhaps selection during domestication was for gene family reduction to eliminate nonessential genes, or perhaps neutral forces were involved.

Gene Evolution in the Relatively Unimproved *Ghp*. Wild tetraploid cotton species naturally occur in habitats that periodically are subjected to drought or salt-stress (7, 17). The relatively unimproved *Ghp* may harbor genes/gene families for adaptation to these harsh environments. Notably, we identified 446 expanded gene families containing 948 genes in *Ghp* compared to the other 7 allotetraploid cotton species. These expanded gene families were significantly associated with the functional terms "sodium ion transport (GO: 0006814, $P = 3.95e-08$)," "glycolytic process (GO: 0006096, $P = 1.7e-04$)," "biotin metabolism (ath00780, $P = 1.20e-05$)," "fatty acid metabolism (ath01212, $P = 1.49e-05$)," and "fatty acid biosynthesis

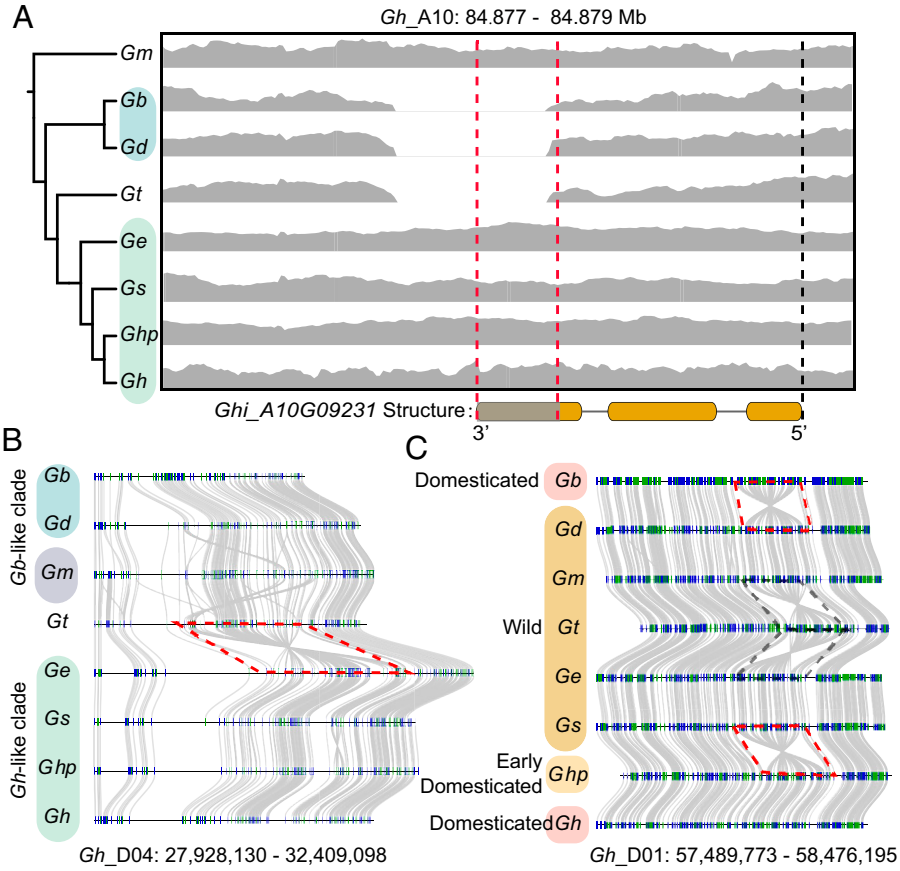

Fig. 4. SVs in tetraploid cotton genomes. (*A*) A 450-bp fragment SV occurred between *Gh*-like and *Gb*-like clades in a synteny block among At10 of all eight tetraploid *Gossypium* genomes. Coverage of *Gh* genome by the Illumina reads of eight tetraploid *Gossypium* genomes (*Upper*) and gene structure of *Ghi_A10G09231* are shown (*Lower*). The deletion region of *Ghi_A10G09231* is outlined in red and marked in gray bar. The yellow bars indicate the coding sequence regions. Evolutionary relationships are shown in the tree to the *Left*. (*B*) A 4.5-Mb inversion occurred between *Gh*-like and *Gb*-like clades. (*C*) A 980-kb inversion occurs at both *Gb* and *Gh* relative to their wild progenitors.

(ath00061, $P = 6.12e\text{-}05$)" (*SI Appendix*, Figs. S15 and S16). To explore possible relationships to stress, we performed transcriptomic sequencing under two independent stress treatments (salt and drought) at three time periods (0, 12, and 24 h) for *Ghp* (*SI Appendix*, Table S18). We detected a total of 9,700 and 1,197 differentially expressed genes (DEGs) in salt and drought treatments, respectively, using 0-h seedling leaves as the control group (Fig. 6*A*). Interestingly, we found that 402 of 948 (42.41%) expanded genes in *Ghp* presented evident transcriptional expression changes, suggesting that expansion of these gene families may have contributed to abiotic and biotic stress resistance. Among these expanded DEGs, we identified one DEG, *GhirPD0101G028900*, homologous to *ECI3* in *A. thaliana*, encoding a homolog of enoyl-CoA δ isomerase 3. This gene is involved in salt and drought stress response in *A. thaliana* (56), and the expression levels of cotton homologs of this gene were also significantly changed under both salt and drought stress treatment in *Ghp* (*SI Appendix*, Figs. S17 and S18). We confirmed that expression of *ECI3* was significantly decreased in the early stages of either cold or salt stress (Fig. 6 *B*–*F*). We also observed a DEG, *GhirPA0801G001500*, a homolog for ethylene-responsive transcription factor RAP2-7, that belongs to the AP2/ERF transcript factor family. This family is not only involved in the regulation of gene expression by stress factors and by components of stress signal transduction pathways, but also negatively regulates the transition to flowering (57). These results exemplify the potential of the relatively unimproved *Ghp* presented here for gene discovery, including those involved in adaptation to environmental challenges (27, 58).

Discussion

Assembling closely related wild and semiwild *Gossypium* genomes has the potential to inform our understanding of the evolution of domesticated *Gossypium* and of adaptation during species diversification. We present three new tetraploid *Gossypium* genome assemblies, including one from an early form of domesticated *G. hirsutum* (race *punctatum*) [$(AD)_1$, *Ghp*], and two recently described wild species of tetraploid cotton, *G. ekmanianum* [$(AD)_6$, *Ge*] and *G. stephensii* [$(AD)_7$, *Gs*]. All three of these have close evolutionary proximity to domesticated *Gh* and share a similar genome size with previously sequenced upland cotton (59–61). Following its initial domestication, *Gh* spread throughout the drier regions of meso-America, in the process developing a diversity of morphological forms spanning the wild-to-domesticated continuum (32). *Ghp* is thought to be among the earliest domesticated forms of *Gh* (32). Our results confirm that they were domesticated from a common wild ancestor. The cryptic *Gh*-like wild species *Ge* and *Gs* are geographically isolated as island endemics in the Caribbean and South Pacific, respectively; however, they are so similar to *Gh* that they have been included as *Gh* in germplasm banks (8, 9). We found that they shared the greatest genetic similarity with *Gh* and *Ghp* and diverged from their most recent common ancestor about 0.75 Mya. We confirm the monophyletic origin of the polyploid clade (35). Notably, *Gh* shares with *Ghp*, *Ge*, and *Gs* a lower level of SNPs and InDel density (59) than found in the other allotetraploid genomes. The three assembled tetraploid *Gossypium* genomes represent a valuable resource for understanding cotton evolution and domestication.

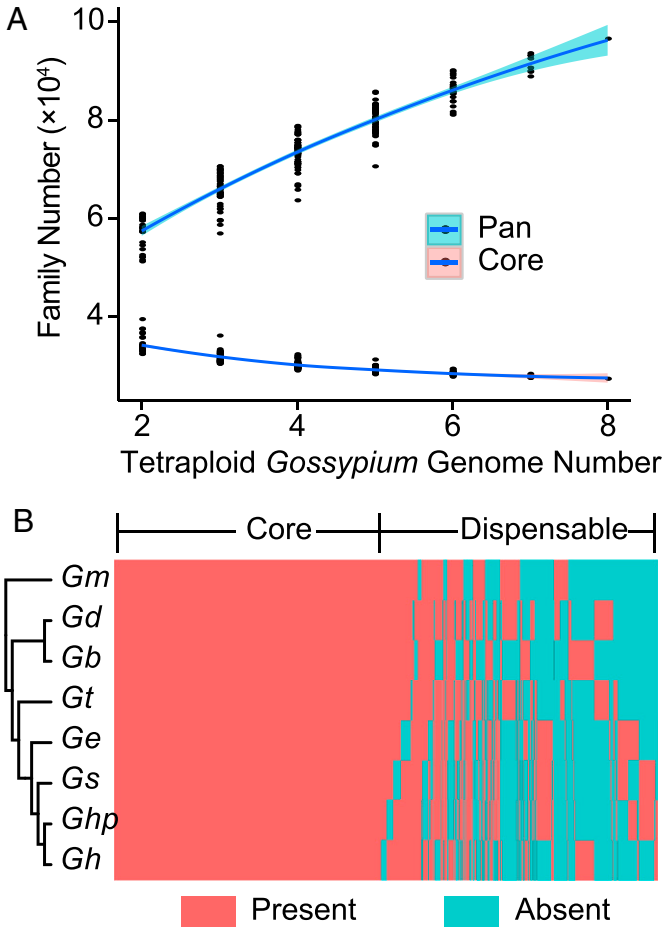

Fig. 5. Pan-genome analysis for eight tetraploid cotton genomes. (*A*) Increase in pan-gene families and decrease in core gene families with the addition of tetraploid cotton genomes. (*B*) Clustering of core and dispensable gene families of tetraploid cotton genomes.

SVs are important in gene expression and functional evolution, and thus are thought to be important in crop domestication and genomic diversity (62–65). Several large chromosomal SVs in domesticated *Gh* have been reported that are related to geographic and genetic differentiation (13, 66, 67). Through identifying SVs among the eight tetraploid *Gossypium* genomes, we affirm the influence of SVs and discovered new SV-related regulatory mutations for specific genes during *Gossypium* evolution. We found a 450-bp SV on At10 that affects fiber length (43), which occurred after divergence from *Gm* and before *Gh*- and *Gb*-like divergence (Fig. 4*A*). Another ∼4.48-Mb inversion on Dt04 accompanies the formation of *Gh*-like species (Fig. 4*B*). These SVs may have contributed to genetic isolation between *Gh*- and *Gb*-like species. One notable SV is an inversion covering 986.42 kb on Dt01, which was detected in both domesticated *Gh* and *Gb* (Fig. 4*C*). This parallelism suggests either a remarkable genomic convergence under human selection, or perhaps more likely, a region of introgression acquired during human manipulation of the two species (21). Collectively, it seems likely that structural variation is a significant factor in *Gossypium* divergence, and that it may have adaptive and agronomic relevance to crop phenotypes.

Crop wild relatives are an important source of agricultural genetic diversity. The observation that the *Gossypium* pan-gene size does not rapidly asymptote suggests that there has been abundant genic diversification that accompanied diversification and global spread of the ∼50 species in the genus (15). The *Gossypium* pan-genome shows the expansion of new genes in *Gh*-like species that can provide the basis for crop improvement. The domestication and crop improvement history of cotton has entailed sequential genetic bottlenecks and an accompanying loss of genetic diversity (68–70); some of this lost diversity might have potential to mitigate problems associated with adaptation to various abiotic and biotic stresses. Compared with domesticated cotton, wild relatives have greater resistance to different abiotic and biotic stresses, including disease, drought, and salinity (7, 17). The relatively unimproved *Ghp* may be useful in this respect, for introgression of favorable genes into domesticated cotton and perhaps to contribute increased resilience to climate challenges.

The completion of reference grade genome sequences for all seven allotetraploid cotton species provides a foundation for future investigation of genomic and phenotypic evolution and adaptation, both in natural settings and under domestication. Future comparative analysis across the full spectrum of "omics" will facilitate a better understanding of cotton evolution and domestication, and further elucidate the genome-level basis of elite traits, such as tolerance to environmental stresses and enhanced cotton fiber properties.

Materials and Methods

Plant Materials and Growth Conditions. Leaves for DNA sequencing were collected from three cotton species, *G. ekmanianum* accession no. *AD602*, *G. stephensii* accession no. *AD701*, and *G. hirsutum* race *punctatum* accession no. *Punctatum 25* (TX-1000). These perennials are all maintained at the National Wild Cotton Nursery in Sanya, China, which is supervised by the Institute of Cotton Research Institute, Chinese Academy of Agricultural Sciences (ICR-CAAS).

Genome Sequencing. High-molecular weight genome DNA (gDNA) of three tetraploid cotton species/accessions (*AD602*, *AD607*, and *Punctatum 25*) was extracted according to the standard CTAB protocol, and subsequently fragmented for PacBio SMRTbell long-read sequencing libraries using Covaris g-TUBE Shearing Device. DNA fragments were purified using 0.45X AMPure beads, and DNA quality was assessed by both Qubit fluorometer and Agilent 2100 Bioanalyzer. The PacBio library was prepared by using the purified DNA fragments and sequenced on the PacBio Sequel I platform.

Illumina paired-end sequencing libraries with an insert size of 350 bp were generated from the same gDNA extraction following the manufacturer's protocol, and all libraries of three species/accessions were sequenced on the Illumina HiSeq X Ten platform as PE150. Illumina Hi-C was generated following a published protocol (71). Briefly, the leaves of 15-d-old seedlings were fixed in 1% formaldehyde solution. The nuclei/chromatin was extracted from the fixed tissue and digested with DpnII. The overhangs resulting from DpnII digestion were filled in using biotin-14-dCTP (Invitrogen) and Klenow (New England Biolabs). After dilution and relegation chromatin with T4 DNA ligase (New England Biolabs), gDNA was extracted and sheared to a size of 300 to 500 bp with a Bioruptor (Diagenode). The biotin-labeled DNA fragments were enriched using streptavidin beads (Invitrogen) and subject to library preparation according to a previous report (72). Illumina sequencers (Illumina HiSeq X Ten platform) carried out the sequencing of the Hi-C libraries. HiC-Pro (v2.10.0) was used to evaluate Hi-C data quality (72). Samples from leaves, stems, and stem apices of mature *Ghp*, *Gs*, and *Ge* plants were collected for extracting RNA. RNA-sequencing (RNA-seq) libraries were constructed using the protocol of NEBNext Ultra RNA Library Prep Kit for Illumina (New England Biolabs) and sequenced on the Illumina X Ten platform.

Contig Assembly. De novo genome assembly was performed mainly using the PacBio SMART long reads with FALCON (https://github.com/PacificBiosciences/FALCON/, falcon-kit==1.8.1) (22). Briefly, we first selected the longest 50× of subreads as seeds to do error correction. These filtered data were used in FALCON for assembly with the parameters: length_cutoff_pr = 5000, max_diff = 100, max_cov = 100. The resulting primary contigs (p-contigs) were then polished using Quiver (https://www.pacb.com/support/software-downloads) (73) by aligning

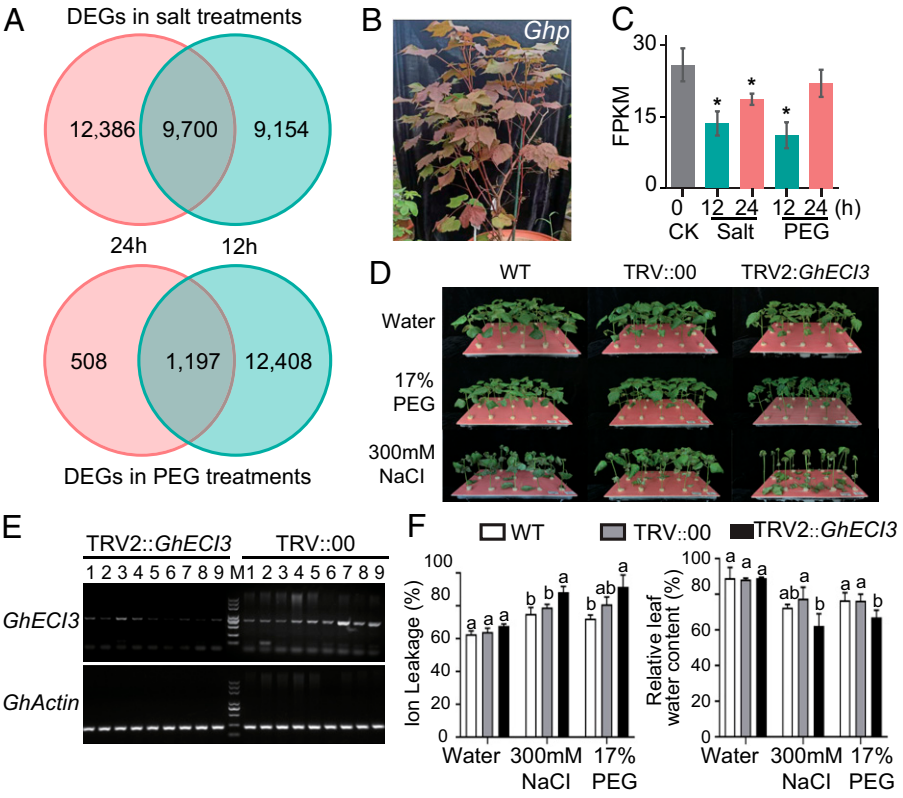

Fig. 6. Abiotic stress adaption of the *Ghp*. (*A*) Venn map of DEGs in salt (300 mM NaCl) and drought (17% PEG) treatments at two time periods, respectively. (*B*) The mature plants of *GhP*. (*C*) Significant comparative analysis of the expression level for *GhECI3* among different treatments. **P* < 0.05. (*D*) Phenotypes of two treatment groups and a control (water) after salt and drought treatments in VIGS plants. (*E*) Expression detection of gene *GhECI3* in its silenced plants (TRV2::*GhECI*). (*F*) Significant comparison analysis of ion leakage and relative leaf water content. ANOVA analysis was performed with the standard *t* test, with least significant difference used for multiple comparisons. The different letters above the error bars indicate significant differences (*P* < 0.05) in all combinations.

total SMRT reads. Finally, Pilon (v1.18) (74) were used to perform a second round of error correction with Illumina PE reads (insertion size = 350 bp).

Chromosome Assembly Using Hi-C. To avoid artificial bias, the following type of reads were removed: 1) reads with ≥10% unidentified nucleotides (N); 2) reads with >10 nt aligned to the adapter, allowing ≤10% mismatches; 3) reads with >50% bases having phred quality < 5. The filtered Hi-C reads were aligned against the contig assemblies with BWA (v0.7.8) (75). Reads were excluded from subsequent analysis if they did not align within 500 bp of a restriction site or did not uniquely map, and the number of Hi-C read pairs linking each pair of scaffolds was tabulated. LACHESIS (https://github.com/shendurelab/LACHESIS) (76) used hierarchical agglomerative clustering to 26 groups. Juicebox v1.22 (https://github.com/aidenlab/Juicebox) was finally used to order the scaffolds in each group.

Assembly Assessment. The genome assembly was evaluated by mapping the high-quality reads from 350-bp insert size PE libraries to the Hi-C assembly using BWA-mem. The distribution of the sequencing depth at each position was calculated to measure the completeness of the genome assembly. BUSCO (v3.0.2) (23) was used to evaluate the assembly completeness of three cotton genomes with 1,440 embryophyte genes from the "Embryophyta_odb9" database. LAI was used to evaluate assembly continuity and completeness by full-length LTR retrotransposons (LTR-RTs) (24). LTRharvest (v1.5.3) (77) (parameters: "-similar 85.00 -vic 10 -seed 30 -seqids yes -motif TGCA -motifmis 1 -minlenltr 100 -maxlenltr 3500 - mindistltr 1000 -maxdistltr 20000 -mintsd 4 -maxtsd 20") and LTR_FINDER (V 64–1.0.5) (78) (parameters: "-w 2 -l 100 -L 3500 -d 1000 -D 20000 -M 0.3") were used to de novo predict the candidate LTR-RTs in the three genome assemblies. LTR_retriever (v2.9.0) (79) was then used to combine and refine all the candidates to get the final, complete LTR-RTs. Each LAI score was calculated based on the formula: LAI = (intact LTR-RTs length/total LTR-RTs length) × 100%.

Repeat Annotation. Repeat annotation was carried out based on de novo predictions and homolog-based predictions for the three new cotton genomes. For de novo-based predictions, RepeatModeler1 (v1.0.8), RepeatScout (v1.0.5), and LTR_FINDER (v1.07) were used to predict TEs and to build a TE library. We integrated this TE library with a known repeat library (Repbase v15.02, homolog-based) and used these with RepeatMasker (v3.3.0) to predict TEs. RepeatProteinMask (v3.3.0, www.repeatmasker.org/RepeatMasker) which makes homology-based predictions, was performed to detect TEs in these three cotton genomes by comparing them to the TE protein database. Tandem repeats were detected in the genome using Tandem Repeats Finder (TRF, v4.07b).

Gene Annotation. A combination of de novo, homology-based, and RNA-seq–based predictions were employed to annotate the PCG in the three cotton genomes. Five ab initio gene-prediction programs were used to predict genes, including Augustus (v3.0.2) (80), Genescan (v1.0) (81), Geneid (v1.4) (82), GlimmerHMM (v3.0.2) (83), and SNAP (v2013-02-16) (84). Protein sequences from six dicot species [i.e., *A. thaliana* (85), *Theobroma cacao* (86), *Populus trichocarpa* (87), *G. hirsutum* (14), *G. arboreum* (14), and *G. raimondii* (29)] were downloaded from CottonGen (88), Ensembl (89), and the National Center for Biotechnology Information (NCBI) (90) and aligned against to the genome using WUblast (v2.0) (91). Genewise (v2.2.0) (92) was employed to predict gene models based on the sequence alignment results. For RNA-seq–based predictions, reads from more than four tissues (leaves, stems, and stem apices) RNA-seq data, which were detected in our research were, aligned to the three cotton genomes using TopHat (v2.0.13) (93) to identify exons region and splice positions. The alignment results were then used as input for cufflinks (v2.1.1) (94) to assemble transcripts to the gene models. In addition, the RNA-seq data were assembled by Trinity (v2.1.1), creating several pseudo-ESTs, which were mapped to each assembly by BLAT (v3.2.3) (95) and used to predict gene models via PASA (r20140417) (96). A weighted and nonredundant gene set was generated by EVidenceModeler (EVM, v1.1.1) (97), which merged all genes models

predicted by the above three approaches. These were combined with the transcript assembly, and PASA was used to adjust the gene models generated by EVM.

Resistance Gene Analog Identification and Evolution. To predict resistance gene analogs (RGAs) in cotton tetraploids, RGAdb from RGAugury (https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-3197-x) software was downloaded (98). Protein sequences of all annotated genes of cottons were aligned to the RGAdb using BLASTP with an E-value cutoff of 1e-05. Seven RGA-related domains and motifs–including NB-ARC, NBS, LRR, TM, STTK, LysM, CC, and TIR–were searched by InterProScan, hmmscan, and phobius from RGAugury pipeline in annotated genes.

Functional Annotation. Predicted protein sequences were assigned functions by searching six protein/function databases: NR, InterPro, GO, Kyoto Encyclopedia of Genes and Genomes (KEGG), Swiss-Prot, and TrEMBL. We used InterproScan46 (v20180213) (99) to search the InterPro database with parameters: -f TSV -dp -gotermes -iprlookup -pa. For the other five databases, BLAST was run with an E-value cutoff of 1e-5. Results from these databases were concatenated together. The R package Clusterprofiler47 (100) was used for GO term and KEGG enrichment analyses.

Orthology and Pan-Genome Analysis. Protein sequences of annotated genes from eight genomes [*G. kirkii* (30), *G. australe* (27), *G. longicalyx* (26), *G. herbaceum* (14), *G. arboreum* (25), *G. thurberi* (28), *G. turneri* (29) and *G. raimondii* (29)] were analyzed in conjunction with the proteins from the eight genomes from allotetraploids (*Gh* (15), *Gb* (15), *Gt* (15), *Gm* (15), *Gd* (15), *Ge*, *Gs*, and *Ghp*) to determine orthology. The longest proteins for each gene were in an all-versus-all BLASTP with an E-value cutoff of 1e-5. OrthoFinder (v2.2.7) (101) was used to detect orthogroups of homologous genes from all genomes using default parameters. Single-copy gene orthogroups were aligned with MUSCLE (v3.8.31) (102) and concatenated into a superalignment. RAxML (v8.0.19) was used to build a phylogenetic tree with the parameters: "-n cds -m GTRGAMMA -p 12345 -x 12345 -# 1000 -f ad". K_a and *Ks* values were calculated for single-copy orthologous genes between each diploid cotton genome and each tetraploid cotton subgenome by KaKs_Calculator (v2.0) software (103). Divergence times (T) were estimated using the formula $T = Ks/2r$ (substitution rate $r = 2.6 \times 10^{-9}$) (104).

Putative positively selected genes were detected using the branch-site model in PAML (v4.7) (105). Genome synteny blocks containing at least four genes was detected using mcscan (https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version)) with parameter: –cscore = 0.90, –iter = 1. Gene families for the eight tetraploid cottons (*Gh*, *Gb*, *Gt*, *Gm*, *Gd*, *Ge*, *Gs*, and *Ghp*) were generated by OrthoFinder. Gene families that were shared among the eight genomes were defined as core gene families, and those that only existed in one genome were defined as species-special gene families. The gene families that were presenting in one to seven samples were defined as dispensable gene families.

Virus-Induced Gene Silencing. Virus-induced gene silencing (VIGS) of the *GhECI3* was performed to verify their potential functions. Here, we used *G. hirsutum* race *Marie-Galante 85* since it has been demonstrated to have better salt stress tolerance in our previous study (106). First, the VIGS vector TRV::*GhECI3* was constructed by recombining ~300-bp fragments of *GhECI3* into pTRV-RNA2 vector and introducing into *Agrobacterium tumefaciens* strain GV4104. TRV::00, without recombined fragments, was used as a control vector. Then, this *Agrobacterium* culture was used to infect seedlings of *G. hirsutum* race *Marie Galante 85* (MAR85) according to a previous protocol (27). The transformed cotton seedlings were grown under greenhouse conditions with 25 °C and 8-h dark/16-h day cycle. After 20 d post-*Agrobacterium* inoculation, the VIGS-plants and non-VIGS plants were exposed to salt (300 mM NaCl) and drought (17% PEG6000) treatment for 3 d. Finally, we collected the leaves of TRV::*GhECI3*, TRV::00, and non-VIGS seedlings for morphological and physiological analysis.

Oligo Probes. Six probes for RGAs were designed based on the *Ge* genome sequence. These oligo probes were synthesized by Ningbo Kangbei Biochem, which attached a 6-carboxyfluorescein (6-FAM) or 6-carboxytetramethylrhodamine (TAMRA) to the 5′ end. Primer sequence information is shown in *SI Appendix*, Table S19. The oligo probes were designed according to a previous method (107). Briefly, the RGA sequences enriched in the chromosomes A04, A11, and D11 of *Ge* were analyzed using the TRF algorithm, using alignment parameters of 2, 7, and 7 for match, mismatch, and indels, respectively. The tandem repeats in each chromosome were identified based on a minimum alignment score of 50 and were divided into three classes with different size of period distances (<20, 20 to 60, and >60). At the same time, the tandem repeats were physically mapped onto the genome sequence using a web server B2DSC (mcgb.uestc.edu.cn/b2dsc) to predict the distribution on chromosomes. The RGA repeat sequences specific to these chromosomes in the genome were determined using the SPSS software (v22.0, SPSS).

FISH Analysis of RGA-Derived Oligo Probes. Root tips of five cotton species–*G. hirsutum* (cultivar: TM-1), *G. barbadense* (cultivar: 3-79), *G. tomentosum* (accession no. in ICR-CAAS: AD3-LZ), *G. mustelinum* (accession no. in ICR-CAAS: AD4-LZ), and *G. darwinii* accession (accession no. in ICR-CAAS: AD5-07)–were harvested from circa 6-d-old incubator-grown seedlings. Root tips were pretreated using 0.089 mM cycloheximide at 20 °C for 80 min, fixed in methanol-acetic acid (3:1), and then stored at 4 °C for 24 h. Chromosome preparations of metaphase chromosomes were created according to a previously reported method (108). The protocol of ND-FISH using synthesized probes was described by Tang et al. (109). Briefly, 10 μL of hybrid solution with 1.0 μL working solution of each probe and residual volume of 2× SSC 1× TE (pH7.0) were added to the metaphase chromosome slides of different cotton species and covered with a plastic film cover. Hybridization took place at 42 °C for 1 to 3 h. After hybridization, the slides were replaced in 2× SSC solution until the plastic film cover fell off naturally. Slides were dried in the dark. Chromosomes were counter-stained with DAPI in Vectashield antifading solution (Vector Laboratories) under a coverslip. Slides were examined using Zeiss Imager M2 microscope. FISH images were captured using CCD camera (MetaSystems CoolCube 1). The photos and signals were merged using MetaSystems Isis software.

RNA-Seq. All samples for RNA-seq were collected from the National Wild Cotton Nursery in Sanya, China. Four-week-old seedlings of *Ghp* and *Gh* (TM-1) were exposed to both salt (300 mM NaCl) and polyethylene glycol (200 g/L PEG). Leaf samples were collected post treatment at 0, 12, and 24 h. Experiments were reproducible in three independent repetitions. In the follow-up analysis, we use the samples at 0 h as the control group (CK), and other samples at 12 h and 24 h after salt or PEG treatment as a different treatment group (S12P, S24P, S12R, and S24R, represent 12-h PEG, 24-h PEG, 12-h NaCl, and 24-h NacCl treatments, respectively). All fresh tissues were frozen in liquid nitrogen immediately and stored at −80 °C before processing. Total RNAs for each sample were extracted using TRIzol Reagent (Invitrogen) according to the manufacturer's instructions. RNA-seq libraries were prepared using the Illumina standard mRNA-seq library preparation kit (Illumina) and sequenced on an Illumina NovaSeq platform as pair-end short reads (150 bp).

RNA-seq data were mapped to the corresponding genomes using Tophat2 (v2.0.8) (93). HTSeq v0.6.1 (110) was employed to count the number of reads mapped to each gene. FPKM (fragments per kilobase of transcript per million mapped reads) was calculated for each gene based on the length of the gene and number of reads mapped to that gene. Differential expression analysis of two groups was (treatment group vs. control group) performed using the DESeq R package (1.18.0) (111). Genes with an adjusted $P < 0.05$ were considered differentially expressed.

Genomic SV Detection. We aligned seven allotetraploid cotton genomes to the *Gm* (AD4_JGI) reference genome and then applied three methods to identify SVs, including smartie-SV (https://github.com/zeeev/smartie-sv) (39), SyRI (https://github.com/schneebergerlab/syri) (41), and SVMU (https://github.com/mahulchak/svmu) (40). Specifically, the pipeline of Smartie-SV was performed based on the BLASR (v5.3.2) alignment with default parameters; we extracted alignment pairs from any pair of genomes based on nucmer (v3.23) (–mum –maxgap = 500 –mincluster = 1000) to serve as input for the packages SyRI and SVMU with default parameters (112). Then, we aligned Illumina reads of seven tetraploid cotton genomes to the *Gm* reference genome to identify SVs using Breakdancer (v1.3.6). On the basis of the above pipeline, we obtained four raw sets of SVs. For insertions and deletions, we merged four raw set using package Jasmine (v1.0.11, https://github.com/mkirsche/Jasmine) with the parameters "min_support = 1 max_dist = 100 k_jaccard = 9 min_seq_id = 0.2 spec_len = 30", and identified candidate insertions and deletions supported by

at least two methods (62). For inversions, we also only considered candidates supported by at least two methods by using bedtools (113). Annotation of genomic SVs was performed using the package ANNOVAR (v2019Oct24) (114). Genomic SVs were categorized as being in exonic regions (overlapping with a coding exon), intronic regions (overlapping with an intron), splice sites (within 2 bp of a splicing junction), upstream and downstream regions (within a 1-kb region upstream or downstream from the transcription start site), and intergenic regions.

Data, Materials, and Software Availability. Raw sequencing and transcriptome data for the three newly assembled cotton genomes have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive, https://www.ncbi.nlm.nih.gov/sra (BioProject accession no. PRJNA739494) (115). The genome sequence information is available under accession nos. JAHRBX000000000, JAHRHL000000000, and JAHRHM000000000 (116, 117, 118). The genome assembly raw data and the transcriptomic data are available under accession nos. SRR15018757–SRR15111907 (119, 120) and SRR14970131–SRR14970150 (121, 122), respectively. All other study data are included in the main text and supporting information.

ACKNOWLEDGMENTS. We thank the National Natural Science Foundation of China (31621005, 32171994, 32072023, 31471548), the Central Plains Science and Technology Innovation Leader Project (214200510029), the Program for Innovative Research Team (in Science and Technology) in University of Henan Province (20IRTSTHN021), the Postgraduate Improvement Project of Henan Province (YJS2022JD47), and the National Key R&D Program of China (2021YFE0101200) for financial support. B.Z. acknowledges support from the US National Science Foundation (Award 1658709) and the Cotton Incorporated (21-855 and 15-770). J.F.W. acknowledges the US National Science Foundation Plant Genome Research Program for support. Y.V.d.P. acknowledges funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (833522) and from Ghent University (Methusalem funding, BOF.MET.2021.0005.01).

Author affiliations: [a]Research Base, Anyang Institute of Technology, State Key Laboratory of Cotton Biology, Anyang 455000, China; [b]State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences, Anyang 455000, China; [c]Novogene Bioinformatics Institute, Beijing 100015, China; [d]Ficus Biotechnology, Ankara 06000, Turkey; [e]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; [f]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50011; [g]National Center for Scientific and Technical Research, Rabat 10102, Morocco; [h]Collaborative Innovation Center of Modern Biological Breeding, School of Life Science and Technology, Henan Institute of Science and Technology, Xinxiang 453003, China; [i]Department of Biology, East Carolina University, Greenville, NC 27858; [j]Department of Biotechnology and Bioinformatics, Ghent University, 9052 Ghent, Belgium; [k]Center for Plant Systems Biology, Vlaams Instituut voor Biotechnologie, 9052 Ghent, Belgium; [l]Centre for Microbial Ecology and Genomics, Department of Biochemistry, Genetics, and Microbiology, University of Pretoria, Pretoria 0028, South Africa; and [m]College of Horticulture, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing 210095, China

Author contributions: R.P., S.T., J.L.C., B.Z., M.V.M., Y.V.d.P., and J.F.W. designed research; Y.X., S.T., Z. Zhou, K.W., Y. Wei, Y.L., H.W., G.H., Z. Zhang, C.E.G., Y.H., Yuhong Wang, P.L., T.W., Q.L., Yangyuan Wang, and Q.W. performed research; R.P., Y.X., S.T., T.U., Z.L., Z. Zhou, X.C., and B.Z. analyzed data; and R.P., S.T., T.U., J.L.C., H.G., B.Z., Y.V.d.P., J.F.W., and F.L. wrote the paper.

1. Y. Jiao *et al.*, Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97–100 (2011).
2. L. Comai, The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* **6**, 836–846 (2005).
3. S. P. Otto, The evolutionary consequences of polyploidy. *Cell* **131**, 452–462 (2007).
4. D. E. Soltis *et al.*, Polyploidy and angiosperm diversification. *Am. J. Bot.* **96**, 336–348 (2009).
5. Y. Van de Peer, E. Mizrachi, K. Marchal, The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424 (2017).
6. J. F. Wendel, L. E. Flagel, K. L. Adams, "Jeans, genes, and genomes: Cotton as a model for studying polyploidy" in *Polyploidy and Genome Evolution*, P. S. Soltis, D. E. Soltis, Eds. (Springer, Berlin, 2012), pp. 181–207.
7. J. F. Wendel, C. E. Grover, "Taxonomy and evolution of the cotton genus, *Gossypium*" in *Agronomy Monographs: Cotton*, D. D. Fang, R. G. Percy, Eds. (American Society of Agronomy, Madison, WI, 2015), **vol. 57**, pp. 25–44.
8. J. P. Gallagher, C. E. Grover, K. Rex, M. Moran, J. F. Wendel, A new species of cotton from Wake Atoll, *Gossypium stephensii* (Malvaceae). *Syst. Bot.* **42**, 115–123 (2017).
9. C. E. Grover *et al.*, Molecular confirmation of species status for the allopolyploid cotton species, *Gossypium ekmanianum* Wittmack. *Genet. Resour. Crop Evol.* **62**, 103–114 (2015).
10. A. H. Paterson *et al.*, Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* **492**, 423–427 (2012).
11. K. Wang, J. F. Wendel, J. Hua, Designations for individual genomes and chromosomes in Gossypium. *J. Cotton Res.* **1**, 3 (2018).
12. M. Wang *et al.*, Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. *Nat. Genet.* **51**, 224–229 (2019).
13. Z. Yang *et al.*, Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. *Nat. Commun.* **10**, 2989 (2019).
14. G. Huang *et al.*, 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 (2020).
15. Z. J. Chen *et al.*, Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. *Nat. Genet.* **52**, 525–533 (2020).
16. Y. Hu *et al.*, *Gossypium barbadense* and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton. *Nat. Genet.* **51**, 739–748 (2019).
17. P. A. Fryxell, *The Natural History of the Cotton Tribe (Malvaceae, Tribe Gossypieae)* (Texas A & M University Press, College Station, 1979).
18. J. B. Hutchinson, Intra-specific differentiation in *Gossypium hirsutum*. *Heredity* **5**, 161–193 (1951).
19. D. E. Soltis, C. J. Visger, D. B. Marchant, P. S. Soltis, Polyploidy: Pitfalls and paths to a paradigm. *Am. J. Bot.* **103**, 1146–1166 (2016).
20. Z. J. Chen, Molecular mechanisms of polyploidy and hybrid vigor. *Trends Plant Sci.* **15**, 57–71 (2010).
21. D. Yuan *et al.*, Parallel and intertwining threads of domestication in allopolyploid cotton. *Adv. Sci. (Weinh.)* **8**, 2003634 (2021).
22. C. S. Chin *et al.*, Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
23. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
24. S. Ou, J. Chen, N. Jiang, Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
25. X. Du *et al.*, 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 (2018).
26. C. E. Grover *et al.*, The *Gossypium longicalyx* genome as a resource for cotton breeding and evolution. *G3 (Bethesda)* **10**, 1457–1467 (2020).
27. Y. Cai *et al.*, Genome sequencing of the Australian wild diploid species *Gossypium australe* highlights disease resistance and delayed gland morphogenesis. *Plant Biotechnol. J.* **18**, 814–828 (2020).
28. C. E. Grover *et al.*, Insights into the evolution of the New World diploid cottons (*Gossypium*, Subgenus *Houzingenia*) based on genome sequencing. *Genome Biol. Evol.* **11**, 53–71 (2019).
29. J. A. Udall *et al.*, De novo genome sequence assemblies of *Gossypium raimondii* and *Gossypium turneri*. *G3 (Bethesda)* **9**, 3079–3085 (2019).
30. J. A. Udall *et al.*, The genome sequence of *Gossypioides kirkii* illustrates a descending dysploidy in plants. *Front. Plant Sci.* **10**, 1541 (2019).
31. J. F. Wendel, New World tetraploid cottons contain Old World cytoplasm. *Proc. Natl. Acad. Sci. U.S.A.* **86**, 4132–4136 (1989).
32. J. McD. Stewart, D. M. Oosterhuis, J. J. Heitholt, J. R. Mauney, Eds., *Physiology of Cotton* (Springer, Dordrecht, ed. 1, 2010).
33. C. L. Brubaker, J. F. Wendel, Reevaluating the origin of domesticated cotton (*Gossypium hirsutum*; Malvaceae) using nuclear restriction fragment length polymorphisms (RFLPs). *Am. J. Bot.* **81**, 1309–1326 (1994).
34. J. F. Wendel, R. G. Percy, Allozyme diversity and introgression in the Galapagos Islands endemic *Gossypium darwinii* and its relationship to continental *G. barbadense*. *Biochem. Syst. Ecol.* **18**, 517–528 (1990).
35. C. E. Grover, K. K. Grupp, R. J. Wanzek, J. F. Wendel, Assessing the monophyly of polyploid *Gossypium* species. *Plant Syst. Evol.* **298**, 1177–1183 (2012).
36. A. A. Golicz, J. Batley, D. Edwards, Towards plant pangenomics. *Plant Biotechnol. J.* **14**, 1099–1105 (2016).
37. M. B. Hufford *et al.*, De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* **373**, 655–662 (2021).
38. P. H. Sudmant *et al.*; 1000 Genomes Project Consortium, An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
39. Z. N. Kronenberg *et al.*, High-resolution comparative analysis of great ape genomes. *Science* **360**, eaar6343 (2018).
40. M. Chakraborty, J. J. Emerson, S. J. Macdonald, A. D. Long, Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat. Commun.* **10**, 4872 (2019).
41. M. Goel, H. Sun, W.-B. Jiao, K. Schneeberger, SyRI: Finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
42. K. Chen *et al.*, BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6**, 677–681 (2009).
43. Y. Zhou *et al.*, Cotton (*Gossypium hirsutum*) 14-3-3 proteins participate in regulation of fibre initiation and elongation by modulating brassinosteroid signalling. *Plant Biotechnol. J.* **13**, 269–280 (2015).
44. J. A. Jarillo, A. Leyva, J. Salinas, J. M. Martínez-Zapater, Low temperature induces the accumulation of alcohol dehydrogenase mRNA in *Arabidopsis thaliana*, a chilling-tolerant plant. *Plant Physiol.* **101**, 833–837 (1993).
45. G. L. de Bruxelles, W. J. Peacock, E. S. Dennis, R. Dolferus, Abscisic acid induces the alcohol dehydrogenase gene in *Arabidopsis*. *Plant Physiol.* **111**, 381–391 (1996).
46. F. U. Hoeren, R. Dolferus, Y. Wu, W. J. Peacock, E. S. Dennis, Evidence for a role for AtMYB2 in the induction of the *Arabidopsis* alcohol dehydrogenase gene (*ADH1*) by low oxygen. *Genetics* **149**, 479–490 (1998).
47. C. E. M. Grossi, F. Santin, S. A. Quintana, E. Fantino, R. M. Ulloa, Calcium-dependent protein kinase 2 plays a positive role in the salt stress response in potato. *Plant Cell Rep.* **41**, 535–548 (2022).
48. K. P. Ismond, R. Dolferus, M. de Pauw, E. S. Dennis, A. G. Good, Enhanced low oxygen survival in *Arabidopsis* through increased metabolic flux in the fermentative pathway. *Plant Physiol.* **132**, 1292–1302 (2003).

49. T. Goujon *et al.*, *AtBXL1*, a novel higher plant (*Arabidopsis thaliana*) putative beta-xylosidase gene, is involved in secondary cell wall metabolism and plant development. *Plant J.* **33**, 677–690 (2003).
50. X. Li *et al.*, ETR1/RDO3 regulates seed dormancy by relieving the inhibitory effect of the ERF12-TPL complex on *DELAY OF GERMINATION1* expression. *Plant Cell* **31**, 832–847 (2019).
51. Y. Liu *et al.*, Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162–176.e13 (2020).
52. S. P. Gordon *et al.*, Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat. Commun.* **8**, 2184 (2017).
53. Q. Zhao *et al.*, Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* **50**, 278–284 (2018).
54. H. W. Jung *et al.*, Pathogen-associated molecular pattern-triggered immunity involves proteolytic degradation of core nonsense-mediated mRNA decay factors during the early defense response. *Plant Cell* **32**, 1081–1101 (2020).
55. E. J. Andersen, S. Ali, E. Byamukama, Y. Yen, M. P. Nepal, Disease resistance mechanisms in plants. *Genes (Basel)* **9**, 339 (2018).
56. S. Goepfert *et al.*, Peroxisomal $\Delta(^3),\Delta(^2)$-enoyl CoA isomerases and evolution of cytosolic paralogues in embryophytes. *Plant J.* **56**, 728–742 (2008).
57. M. J. Aukerman, H. Sakai, Regulation of flowering time and floral organ identity by a MicroRNA and its *APETALA2*-like target genes. *Plant Cell* **15**, 2730–2741 (2003).
58. L. Zeng *et al.*, Whole genomes and transcriptomes reveal adaptation and domestication of pistachio. *Genome Biol.* **20**, 79 (2019).
59. T. Zhang *et al.*, Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat. Biotechnol.* **33**, 531–537 (2015).
60. R. Peng, D. C. Jones, F. Liu, B. Zhang, From sequencing to genome editing for cotton improvement. *Trends Biotechnol.* **39**, 221–224 (2021).
61. Z. Yang, G. Qanmber, Z. Wang, Z. Yang, F. Li, *Gossypium* genomics: Trends, scope, and utilization for cotton improvement. *Trends Plant Sci.* **25**, 488–500 (2020).
62. M. Alonge *et al.*, Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**, 145–161.e23 (2020).
63. C. Chiang *et al.*; GTEx Consortium, The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).
64. J. C. Stein *et al.*, Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nat. Genet.* **50**, 285–296 (2018).
65. J.-M. Song *et al.*, Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nat. Plants* **6**, 34–45 (2020).
66. S. He *et al.*, The genomic basis of geographic differentiation and fiber improvement in cultivated cotton. *Nat. Genet.* **53**, 916–924 (2021).
67. Z. Ma *et al.*, High-quality genome assembly and resequencing of modern cotton cultivars provide resources for crop improvement. *Nat. Genet.* **53**, 1385–1391 (2021).
68. D. L. Van Tassel *et al.*, Re-imagining crop domestication in the era of high throughput phenomics. *Curr. Opin. Plant Biol.* **65**, 102150 (2022).
69. S. A. Flint-Garcia, Genetics and consequences of crop domestication. *J. Agric. Food Chem.* **61**, 8267–8276 (2013).
70. B. T. Moyers, P. L. Morrell, J. K. McKay, Genetic costs of domestication and improvement. *J. Hered.* **109**, 103–116 (2018).
71. N. L. van Berkum *et al.*, Hi-C: A method to study the three-dimensional architecture of genomes. *J. Vis. Exp.* **39**, e1869 (2010).
72. N. Servant *et al.*, HiC-Pro: An optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
73. C.-S. Chin *et al.*, Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
74. B. J. Walker *et al.*, Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
75. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv* [Preprint] (2013). https://arxiv.org/abs/1303.3997 (Accessed 15 January 2020).
76. J. N. Burton *et al.*, Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
77. D. Ellinghaus, S. Kurtz, U. Willhoeft, *LTRharvest*, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
78. Z. Xu, H. Wang, LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
79. S. Ou, N. Jiang, LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
80. M. Stanke *et al.*, AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
81. C. Burge, S. Karlin, Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
82. R. Guigó, Assembling genes from predicted exons in linear time with dynamic programming. *J. Comput. Biol.* **5**, 681–702 (1998).
83. W. H. Majoros, M. Pertea, S. L. Salzberg, TigrScan and GlimmerHMM: Two open source *ab initio* eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
84. I. Korf, Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
85. T. P. Michael *et al.*, High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. *Nat. Commun.* **9**, 541 (2018).
86. X. Argout *et al.*, The cacao Criollo genome v2.0: An improved version of the genome for genetic and functional genomic studies. *BMC Genomics* **18**, 730 (2017).
87. B. T. Hofmeister *et al.*, A genome assembly and the somatic genetic and epigenetic mutation rate in a wild long-lived perennial *Populus trichocarpa*. *Genome Biol.* **21**, 259 (2020).
88. J. Yu *et al.*, CottonGen: A genomics, genetics and breeding database for cotton research. *Nucleic Acids Res.* **42**, D1229–D1236 (2014).
89. K. L. Howe *et al.*, Ensembl 2021. *Nucleic Acids Res.* **49** (D1), D884–D891 (2021).
90. P. A. Kitts *et al.*, Assembly: A resource for assembled genomes at NCBI. *Nucleic Acids Res.* **44** (D1), D73–D80 (2016).
91. R. She, J. S.-C. Chu, K. Wang, J. Pei, N. Chen, GenBlastA: Enabling BLAST to identify homologous gene sequences. *Genome Res.* **19**, 143–149 (2009).
92. E. Birney, M. Clamp, R. Durbin, Genewise and genomewise. *Genome Res.* **14**, 988–995 (2004).
93. D. Kim *et al.*, TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
94. S. Ghosh, C.-K. K. Chan, Analysis of RNA-seq data using TopHat and Cufflinks. *Methods Mol. Biol.* **1374**, 339–361 (2016).
95. W. J. Kent, BLAT–The BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
96. B. J. Haas *et al.*, Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
97. B. J. Haas *et al.*, Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
98. P. Li *et al.*, RGAugury: A pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genomics* **17**, 852 (2016).
99. S. Hunter *et al.*, InterPro: The integrative protein signature database. *Nucleic Acids Res.* **37**, D211–D215 (2009).
100. G. Yu, L.-G. Wang, Y. Han, Q.-Y. He, clusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
101. D. M. Emms, S. Kelly, OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
102. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
103. D. Wang, Y. Zhang, Z. Zhang, J. Zhu, J. Yu, KaKs_Calculator 2.0: A toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics* **8**, 77–80 (2010).
104. C. E. Grover *et al.*, Comparative genomics of an unusual biogeographic disjunction in the cotton tribe (Gossypieae) yields insights into genome downsizing. *Genome Biol. Evol.* **9**, 3328–3344 (2017).
105. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
106. Y. Xu *et al.*, Genetic regulatory networks for salt-alkali stress in *Gossypium hirsutum* with differing morphological characteristics. *BMC Genomics* **21**, 15 (2020).
107. X. Liu *et al.*, Dual-color oligo-FISH can reveal chromosomal variations and evolution in *Oryza* species. *Plant J.* **101**, 112–121 (2020).
108. Y. Liu *et al.*, Chromosome painting based on bulked oligonucleotides in cotton. *Front. Plant Sci.* **11**, 802 (2020).
109. S. Tang *et al.*, Developing new oligo probes to distinguish specific chromosomal segments and the A, B, D genomes of wheat (*Triticum aestivum* L.) using ND-FISH. *Front. Plant Sci.* **9**, 1104 (2018).
110. S. Anders, P. T. Pyl, W. Huber, HTSeq–A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
111. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
112. S. Kurtz *et al.*, Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
113. A. R. Quinlan, BEDTools: The Swiss-Army tool for genome feature analysis. *Curr. Protoc. Bioinform.* **47**, 11.12.11–11.12.34 (2014).
114. K. Wang, M. Li, H. Hakonarson, ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
115. R. Peng *et al.*, Tetraploid Cotton Genome sequencing and assembly, PRJNA739494. NCBI BioProject. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA739494/. Deposited 16 August 2022.
116. R. Peng *et al.*, Genome JAHRBX000000000. NCBI. https://www.ncbi.nlm.nih.gov/genome/?term=JAHRBX000000000. Accessed 8 September 2022.
117. R. Peng *et al.*, Genome JAHRHL000000000. NCBI. https://www.ncbi.nlm.nih.gov/genome/?term=JAHRHL000000000. Accessed 8 September 2022.
118. R. Peng *et al.*, Genome JAHRHM000000000. NCBI. https://www.ncbi.nlm.nih.gov/genome/?term=JAHRHM000000000. Accessed 8 September 2022.
119. R. Peng *et al.*, SRR15018757. NCBI SRA. https://www.ncbi.nlm.nih.gov/sra/?term=SRR15018757. Accessed 8 September 2022.
120. R. Peng *et al.*, SRR15111907. NCBI SRA. https://www.ncbi.nlm.nih.gov/sra/?term=SRR15111907. Accessed 8 September 2022.
121. R. Peng *et al.*, SRR14970131. NCBI SRA. https://www.ncbi.nlm.nih.gov/sra/?term=SRR14970131. Accessed 8 September 2022.
122. R. Peng *et al.*, SRR14970150. NCBI SRA. https://www.ncbi.nlm.nih.gov/sra/?term=SRR14970150. Accessed 8 September 2022.