Allotetraploid cotton (*Gossypium*) 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 Ge, Gs, 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.

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 accesses 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).
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_obd10) (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, Ge, Gb, Gossypium tomentosum [Gb, (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 telomers, having an average of 0.85-fold (P < 10−16, Wilcoxon test) and 2.34-fold (P < 10−16, Wilcoxon 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, Wilcoxon test) in telomeric versus other regions.

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       |
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 *Gb*-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. 2A): 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 *Gb*-like clade for the two subgenomes [Dt clade: (*Ge, Gs, Ghp*)] vs. At clade: {*Ge, Gs, (Gb, Ghp)*} (Fig. 2A), 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 *A₁* genome of tetraploid cotton is more divergent from *A₂* (~1.31 Mya) than *A₁* (~1.23 Mya), the range in estimates for these divergence times overlap (*SI Appendix, Fig. S3*).
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 Gb- and Gb-like clades. In general, the synonymous substitution rate (Ks) was higher for Dt homeologs than for At homeologs (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 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 Gb) 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 genetic PAV-induced changes across the polyploid phylogeny (0.078) by a factor of

![Fig. 2. Phylogenetic analysis of the Gossypium genomes. (A) Maximum-likelihood tree inferred using G. kirkii (Gk) 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.](image-url)
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. 4A). This shorter version is phylogenetically homoplasic, 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 Gb-like clade from the Gt and the Gb-like clade (Fig. 4B), and thus is phylogenetically diagnosed as having occurred in the ancestor of the Ge through Gb 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 Gb, but not their close wild relatives. At its boundary, the gene Ghi_D01G09866 was the homologous BXL in Arabidopsis thaliana, which encodes a β-xylodase 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 (Gb, Gb, Gm, and Gd) (15), a total of 96,537 gene families was detected (Fig. 5A). 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. 5A), 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. 5B and SI Appendix, Table S16). We also observed 6.81% of genes, on average, were species-specific, but the number and the proportion of Gb-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., Gb 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 Gb-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),” “bacterial metabolism (ath00780, P = 1.20e-05),” “fatty acid metabolism (ath01212, P = 1.49e-05),” and “fatty acid biosynthesis
We conadaptation to environmental challenges (27, 58). The results exemplify the potential of the relatively unimproved but also negatively regulates the transition to factors and by components of stress signal transduction pathways, not only involved in the regulation of gene expression by stress homolog for ethylene-responsive transcription factor RAP2-7, that decreased in the early stages of either cold or salt stress (Fig. 6 results of this gene were also significantly changed under both salt and drought treatments, respectively, using 0-h seedling leaves as the control group (Fig. 6A). 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, GhiRD0101G028900, homologous to EC13 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. S15 and S16). We confirmed that expression of EC13 was significantly decreased in the early stages of either cold or salt stress (Fig. 6B–F). We also observed a DEG, GhiPA0801G001500, 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)2, Ge] and G. stephensii [(AD)3, 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.
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 Gb 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 Gb- and Gb-like divergence (Fig. 4A). Another ∼4.48-Mb inversion on Dt04 accompanies the formation of Gb-like species (Fig. 4B). These SVs may have contributed to genetic isolation between Gb- and Gb-like species. One notable SV is an inversion covering 986.42 kb on Dt01, which was detected in both domesticated Gb and Gb (Fig. 4C). 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 an important 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-genome shows the expansion of new genes in Gb-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 gTUBE 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 filled in using biotin-14-DCP (Invitrogen) and Klenow (New England Biolabs). After dilution and reactivation in 14 DNA ligase (New England Biolabs), gDNA was extracted and sheared to a size of 300 to 500 bp with a Bioruptor (Diagenode). The bait-labeled DNA fragments were enriched using streptavidin beads (Invitrogen) and subject to library preparation according to a previous report (72). Illumina sequences (Illumina HiSeq X Ten platform) carried out the sequencing of the Hi-C libraries. Hi-C 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 Gp 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/PacBioBiosciences/FALCON), falcon-kit=1.8.1) (22). Briefly, we first selected the longest 50x 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.pacbio.com/support/software-downloads) (73) by aligning...
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 (v3.0.2) (23) was used to evaluate the assembly completeness of three cotton genomes with 1,440 embryophyte genes from the NCBI (90) and aligned against the genome using WUblast (v2.0.9) (91). Genewise (v2.2.0) (92) was employed to predict gene models based on de novo 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 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), Ensemble (89), and the National Center for Biotechnology Information (NCBI) (90) and aligned against the genome using WUblast (v2.0.9) (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 (v20130417) (96). A weighted and nonredundant gene set was generated by EVidenceModeler (EVM, v1.1.1) (97), which merged all genes models

**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 LTRFinder (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), Ensemble (89), and the National Center for Biotechnology Information (NCBI) (90) and aligned against the genome using WUblast (v2.0.9) (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 (v20130417) (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 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://bmgenomecs.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, NLR, TRM, STIK, 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 -d gotemres -iplookup -p. 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. australis (27), G. longicalyx (26), G. herbaceum (14), G. arboresum (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 (15), Gs (15), and Gh (15)) to determine orthology. The longest proteins for each gene were in an all-versus-all BLASTP with an E-value cutoff of 1e-05. 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 using default parameters. Single-copy gene orthogroups were aligned with MUSCLE (v3.8.31) (102) and concatenated into a superalignment.

**Divergence times** 

\[ \tau = 2.6 \times 10^{-3} \times 10^{9} \]  

(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/tanghaibo/cvjiki/MI/McscaN (Python version)) with parameters: -c score = 0.90, -iter = 1. Gene families for the eight tetraploid cottons (Gh, Gb, Gt, Gm, Gd, Ge, Gs, and Gh) 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 present in one to seven samples were defined as dispensable gene families.

**Viruses-Induced Gene Silencing.** Virus-induced gene silencing (VIGS) of the GHEC3 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::00, GtRECA (v8.0.19) was used to build a phylogenetic tree with the parameters: MUSCLE (v3.8.31) (102) and concatenated into a superalignment. RAxML using default parameters. Single-copy gene orthogroups were aligned with MUSCLE (v3.8.31) (102) and concatenated into a superalignment.

**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 NaCl 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).

**SVs 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/schneeberger/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) (maxmatch = 500 –minclust = 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/mikirsch/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 FISH Analysis of Gm-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.awinii accession (accession no. in ICR-CAAS: AD5–G7)—were harvested from circa 6-d-old inoculator-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 (pH 7.0) 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 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 cover-slip. 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.
at least two methods (62). For inversions, we also only considered candidates supported by at least two methods using bedtools (113). Annotation of genomic SVs was performed using the package ANNOVAR (v020409c24) (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. PRJNA739449) (115). The genome sequence information is available under accession nos. JAHRRX000000000, JAHRRH000000000, and JAHRRHM000000000 (116, 117, 118). The genome assembly raw data and the transcriptomic data are available under accession nos. SRR15018757–SRR15111907 (119, 120) and SRR14970151–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 in Henan Province (20RTSHN021), the Postgraduate Improvement Project of Henan Province (YJS2022D47), 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.2021.0001.01).

Author affiliations: *Research Base, Anyang Institute of Technology, State Key Laboratory of Cotton Biology, Anyang 455000, China; **State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences, Anyang 455000, China; ***Novogene Bioinformatics Institute, Beijing 100015, China; ****Ficus Biotechnology, Ankara 06000, Turkey; *****Shenzhen Branch, Guangdong Laboratory for Jiangnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China; ***Department of Ecology, Evolution, and Conservation Biology, Iowa State University, Ames, IA 50011; National Center for Scientific and Technical Research, Rabat 10102, Morocco; ****Collaborative Innovation Center of Modern Tropical Biological Engineering, School of Life Science and Technology, Henan Institute of Science and Technology, Xinxiang 453003, China; Department of Biology, East Carolina University, Greenville, NC 27858; Department of Biotechnology and Bioinformatics, Ghent University, 9027 Ghent, Belgium; **Center for Plant Systems Biology, Vlaams Instituut Voor Biotechnologie, 9052 Ghent, Belgium; Centre for Microbial Ecology and Genomics, Department of Biochemistry, Genetics, and Microbiology, University of Pretoria, Pretoria 0028, South Africa; and ***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, X.L., T.W., Q. Yang, G. Yin, G. Zhang, and N. Wu, performed research; B.Y., 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., F.W., and L.E. wrote the paper.

1. Y. Jiao et al., Ancestral polyplody in seed plants and angiosperms. Nature 473, 97–100 (2011).
2. I. J. Conner, The advantages and disadvantages of being polyplody. 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. X. Van de Peer, E. Muzzio, 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, 2013), pp. 181–203.
7. J. F. Wendel, C. E. Grover, “Taxonomy and evolution of the cotton genen, Gossypium” in Agronomy Monographs: Cotton, D. D. Fang, R. G. Perry, 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 Woke Atoll, Gossypium stephensi (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. Crop Res. 1, 3 (2018).
12. M. Wang et al., Reference genome sequences of two cultivated allotetraploid cottons, Gossypium hirsutum L. and Gossypium arboreum (Malvaceae). 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, 2899 (2019).
14. G. Huang et al., Genome sequence of Gossypium hirsutum 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, 529–533 (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. Freyfельд, The Natural History of the Cotton Tribe (Malvaceae, Tribe Gossypiinae) (Texas A&M University Press, College Station, 1979).
18. J. B. Hutchinson, Intra-specific differentiation in Gossypium hirsutum. Heredity 56, 1–193 (1951).
19. D. E. Soltis, C. J. Viergo, D. B. Marchant, P. S. Soltis, Polyplody: 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., Paralleling and interweaving threads of domestication in allotetraploid cotton. Adv. Sci. (Weinh.) 8, 2003634 (2021).
22. C. E. Chin et al., Phased genome assembly with single-molecule real-time sequencing. Nat. Methods 13, 1050–1054 (2016).
23. A. A. P. Fortin, M. R. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zibinova, BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 33, 3210–3212 (2017).
24. S. Ou, J. Chen, N. Jiang, Assessing genome assembly quality using the LIR Assembly Index (LAI). Nucleic Acids Res. 46, e162 (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 linguis caseulaceum as a resource for cotton breeding and evolution. G3 (Bethesda) 10, 1457–1467 (2020).
