# Comparative genomics and polyploid dynamics in tetraploid cotton (Gossypium)

**Renhai Peng**
Research Base，Anyang Institute of Technology, State Key Laboratory of Cotton Biology
**Yanchao Xu**
State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences
**Zhen Liu**
Research Base，Anyang Institute of Technology, State Key Laboratory of Cotton Biology
**Liyang Chen**
Novogene Bioinformatics Institute https://orcid.org/0000-0002-5639-7305
**Zhongli Zhou**
State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences
**Xiaoyan Cai**
State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences
**Kunbo Wang**
Cotton Research Institute
**Yangyang Wei**
Anyang Institute of Technology
**Yuling Liu**
Anyang Institute of Technology
**Heng Wang**
State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences
**Guan-jing Hu**
Department of Ecology, Evolution and Organismal Biology, Iowa State University
https://orcid.org/0000-0001-8552-7394
**Corrinne Grover**
Iowa State University https://orcid.org/0000-0003-3878-5459
**Yuqing Hou**
State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences
**Yuhong Wang**

State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences

**Pengtao Li**
Anyang Institute of Technology

**Tao Wang**
Anyang Institute of Technology

**Qaunwei Lu**
Anyang Institute of Technology

**Yuanyuan Wang**
Collaborative Innovation Center of Modern Biological Breeding, School of Life Science and Technology, Henan Institute of Science and Technology

**Qinlian Wang**
Collaborative Innovation Center of Modern Biological Breeding, School of Life Science and Technology, Henan Institute of Science and Technology

**Shilin Tian**
Novogene Bioinformatics Institute https://orcid.org/0000-0001-8958-1806

**Jonathan Wendel**
Iowa State University https://orcid.org/0000-0003-2258-5081

**Fang Liu** ( **liufcri@163.com** )
State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences https://orcid.org/0000-0002-6213-9572

---

**Article**

**Keywords:** cotton, Gossypium, polyploids, plant evolution

**Posted Date:** July 24th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-734586/v1



---

# Abstract

Allotetraploid species of cotton (*Gossypium*) represent a model system for the study of plant polyploidy, molecular evolution and domestication. In this study three high-quality draft assemblies of tetraploid cottons are presented, comprising one early form of domesticated *Gossypium hirsutum* ($AD_1$-genome, *Gh*), i.e., *Gossypium hirsutum* race *punctatum* (*GhP*), and two recently described wild species of tetraploid cotton, *G. ekmanianum* ($AD_6$, *Ge*) and *G. stephensii* ($AD_7$, *Gs*). Using comparative phylogenomics, we confirm a monophyletic origin of tetraploid *Gossypium* and provide a dated whole-genome level perspective for the evolution of the clade. Recombination and patterns of selection are asymmetric between the two co-resident genomes in the allopolyploid nucleus. Considerable gene structural variation occurs widely within homoeologous genomes and between heterologous genomes during evolution and domestication. Despite few large-scale chromosomal structure variations among tetraploid cotton, frequent homoeologous exchanges between subgenomes in all species have contributed to diversity and asymmetrically between subgenomes. Abiotic and biotic adaptive evolution was driven by various evolutionary forces, leading to transcriptome change and gene family expansion. Our study marks a milestone in modern polyploid crop research, completing genome sequencing for all species of polyploid *Gossypium*, and will facilitate a better understanding of the genomic landscape and crop improvement dynamics of polyploids.


# Main

Polyploidization is an important evolutionary process in many higher plants, leading to new lineages and ecological adaptations[1-5]. Allotetraploid has a natural heterosis per se. Approximately 1-2 Million years 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 ($AD_1$ – $AD_7$)[8, 9]. Tetraploid cottons provide an important model system for understanding evolutionary consequences of polyploids, as well as dual domestication[10-12].

Among the seven species of allopolyploid *Gossypium*, two, i.e., *Gh* (*G. hirsutum,* genome designation $AD_1$) and *Gb* (*G. barbadense,* genome designation $AD_2$), provide the majority of natural fiber for commercial production[13, 14]. Five tetraploid cottons ($AD_1$-$AD_5$) recently have been sequenced using long read technology, providing genome assemblies and resources for uncovering the genetic basis of spinnable fiber formation and development[15-19]. To date, genome assemblies have not been generated for the two most recently described wild tetraploid species, both closely related to *Gh*, i.e., *Ge* (*G. ekmanianum,* $AD_6$) from the Dominican Republic and *Gs* (*G. stephensii.* $AD_7$) from the Wake Atoll near French Polynesia. Moreover, no genome sequences have been generated for primitively domesticated or wild forms of either of the two domesticated species. Among the great diversity of morphological forms spanning the wild-to-domesticated continuum in *G. hirsutum*, 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*).

Genome structural variations (SVs) often impact gene expression and phenotypes in flowering plants[20-23]. In cotton, a few large-scale chromosomal inversions are associated with population differentiation and environmental adaptation among *Gossypium hirsutum* cultivars, consistent with the observation that SVs may drive differentiation[16]. In addition to SVs, homoeologous exchange (HE) between subgenomes may affect chromosome balance[24] and create diversity[25, 26]. Although HEs are thought to be uncommon in cotton, unequal homoeologous exchanges of repeat sequence between A and D subgenomes subsequent to their formation might explain some of the changes in A and D subgenome size that arose after polyploidization[18, 27]. Here, we report high-quality genome assemblies of the three tetraploid genomes, *Ge, Gs,* and *GhP*. Using comparative phylogenomics, we reveal extensive SV and HE in tetraploid cottons, and reevaluate phylogenetic relationships and divergence times within the polyploid clade. Extensive structure variations were associated with phenotypic diversity, including the economically important trait fiber length. Unequal HE events between At and Dt subgenomes were observed in all species of tetraploid cotton. These results deepen our understanding of the genetic and morphotype diversities among tetraploid cottons.

# Results

### Genome assembly and comparative genomic analysis

Three previously unsequenced allotetraploid cotton genomes (*Ge*, *Gs,* and *GhP*) were 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[28] using an average of 22.34 million PacBio long reads each (**Supplementary Tables 1 and 2**), 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; **Supplementary Table 3**) were combined with Hi-C interaction information to produce chromosome-scale scaffolds (**Supplementary Tables 4, 5 and Supplementary Fig. 1**), yielding final assemblies of 2.34 Gb, 2.29 Gb 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 over 99% of bases anchored onto chromosomes and where over 99% of mapped Illumina reads covered more than 97% of the genome (**Supplementary Table 6**). Nearly all of the 1,440 Embryophyta Benchmarking Universal Single-Copy Orthologs (BUSCOs)[29] were complete in the *Ge* (95.5%), *Gs* (97.1%), and *GhP* (95.4%) assemblies (**Supplementary Table 7**), and 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'[30] (**Supplementary Table 8**).

A total of 1,575 Mb (65%), 1,489 Mb (63%) and 1,488 Mb (65%) of transposable elements (TEs) were predicted in *Ge*, *Gs,* and *GhP*, respectively (**Table 1 and Supplementary Table 9**). By combining both homology- and *ab initio*-based methods with transcriptional information (**Supplementary Table 10**), 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 (**Supplementary Table 11**). Most (95 – 97%) of PCGs predicted in *Ge*, *Gs,* and *GhP* had an identifiable homolog (>80% protein identity) in the published tetraploid cotton genomes (**Supplementary Table 12**), i.e., *Gh*[18], *Gb*[18], *Gt* (*G. tomentosum,* $AD_3$)[18], *Gm* (*G. mustelinum,* $AD_4$)[18], and *Gd* (*G. darwinii,* $AD_5$)[18]. An assessment of TE and PCG density in 1000 sliding 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 other chromosomal regions (**Supplementary Figs. 2 and 3**). In contrast, *Gypsy* element density exhibited an average decrease of 0.74-fold ($P < 10^{-16}$, Wilcox test) in telomere versus other regions.

We generated an initial assessment of the tetraploid cotton pangenome by combining our newly sequenced genomes with the five previously published tetraploid cotton species (*Gh*, *Gb*, *Gt*, *Gm* and *Gd*) [18]. As expected, the number of “dispensable” genes increased as additional genomes were added, whereas the number of core genes decreased (**Fig. 1a).** We found that the *Gossypium* tetraploid pangenome was composed of 37,846 gene families, most of which (72.6%, or 27,483 families) were considered core families that account for an average of 68% of the genes in each genome. Approximately one-quarter of the gene families were considered dispensable in each genome (an average of 13,437 families), and extremely few (2-12 families) were considered species-specific (**Fig. 1b and Supplementary Table 13**). Core gene families were enriched for gene ontology terms related to “regulation of biosynthetic process” and “metabolic process” (**Supplementary Fig. 4**). Annotation differences notwithstanding, these results suggest that of the availability of more high-quality genomes will promote the understanding of cotton genome diversity. That is, although a portion of these content differences likely reflect errors in assembly or annotation, detailed analyses of these comparative data may yield insight into both mechanisms of gene loss/gain as well as possible functional consequences of this genic content evolution.

## Phylogenetic analysis of tetraploid Gossypium

An updated phylogeny including 17 diploid and polyploid cotton species/accessions was generated using 3,281 single-copy coding genes. The species sampled included the eight tetraploid cottons evaluated here, as well as eight diploid cottons (*G. herbaceum* $A_{1[17]}$, *G. arboreum* $A_{2[31]}$, *G. longicalyx* $F_{1[32]}$, *G. australe* $G_{2[33]}$, *G. thurberi* $D_{1[34]}$, *G. raimondii* $D_{5[13]}$ and *G. turneri* $D_{10[35]}$) and the phylogenetic outgroup species *Gossypioides kirkii* (*Gki*)[36]. A maximum likelihood phylogenetic tree was inferred and divergence times were estimated using *Ks* values for orthologous genes (**Figs. 2a and 2b**). As shown, the phylogeny for allopolyploid cotton is reiterated in both the A and D genome clades, as expected given their formation from A and D genome diploid antecedents. Tetraploid clade divergence time was estimated at 1.80 Mya (95% CI: 1.10 - 2.72 Mya) (**Fig. 2b**), consistent with previous reports[19, 37]. Most of the tetraploid species fall into two clearly distinguished clades, each of which includes one of the two economically important cultivated cottons (upland cotton *Gh* and Sea Island cotton *Gb*), as in

earlier reports. These two groups are hereafter referred to as the *Gh*-like and *Gb*-like clades, respectively (**Fig. 2c**), and are inferred to have diverged ~ 0.79 Mya (95% CI, 0.49 - 1.49 Mya). The *GhP* genome represents an unimproved landrace of upland cotton and is among the most primitive of the many forms of semi-domesticated and/or feral derivatives found within diversity generated by the 4,000+ year history of *G hirsutum* domestication[38, 39]. We observed similar divergence times between *Gb-Gd* (0.63 Mya; 95% CI: 0.37- 1.26 Mya) and *GhP-Gh* (0.68 Mya; 95% CI: 0.41- 1.14 Ma), confirming earlier data indicating that the Galapagos Island endemic *G. darwinii* (*Gd*), previously considered to be conspecific with *G. barbadense*, diverged relatively recently from its mainland relative[40]. The genome sequences for the two species *Ge* and *Gs* complete the sampling of wild tetraploid *Gossypium*, and, as expected from prior analyses[10-12], they fall within the *Gh*-like clade, having all diverged from their most recent common ancestor around 0.75 Mya (95% CI, 0.42 - 1.33 Mya) (**Supplementary Fig. 5**). With respect to diploid divergence, two obvious branches are distinguished (**Fig. 2a**), which were named the New World clade (D genome) and the African-Australia-Asian clade ((A, G, F genomes)[6, 7].

The above results support previous inferences regarding the monophyletic origin of allopolyploid cottons. 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 (**Supplementary Fig. 5**). As 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 homoeologs than At (**Supplementary Fig. 6**), consistent with previous reports[18] and possibly reflecting subgenome-specific evolutionary processes, including differences in recombination rates and selective sweeps. Slight phylogenetic inconsistencies were found in divergence order within the *Gh*-like clade for the two subgenomes (Dt clade: ((*Ge*, *Gs*), (*Gh*, *GhP*)) *vs.* At clade: (*Ge*, (*Gs*, (*Gh*, *GhP*))) **(Fig. 2a**), which is reasonable given the rapid divergence exhibited by these species[7-9].

## Genomic structural variations (SVs)

Genomic structural variations occur frequently during plant evolution and domestication, providing a major genetic source of phenotypic diversity[41]. We focused on identifying all SVs ≥50 bp in length within *Gossypium* genomes because these are the least well-characterized genetic variations and are likely to affect gene function[42, 43]. By mapping the seven tetraploid *Gossypium* assembled genomes and their sequencing reads to the reference genome of *G. mustelinum* (*Gm*: AD4_JGI[18]), four methods (smartie-SV[43], SVMU[44], SyRI[45], and Breakdancer[46]) were combined to identify SVs (**Supplementary Fig. 7**) and polarize their directionality (i.e., insertion vs. deletion relative to *Gm*). SVs were only considered when they were consistent identified by at least two methods, resulting in an average of 72,965 insertions (range from 67,885 to 77,756), 63,126 deletions (range from 59,663 to 65,670), and 339 inversions (range from 297 to 410) (**Supplementary Table 14**). The lowest number of SVs was observed in *Gt* (**Supplementary Fig. 8**). Notably, the cultivated polyploids (*Gb* and *Gh*) had the longest average length of SVs among the seven tetraploid cotton genomes surveyed, yet the other five cotton accessions exhibit more PAVs (insertions and deletions) with a size ≥1Kb (**Supplementary Table 15**). Relative to the

large number of species/accession-specific SVs (range from 36,476 to 75,125), fewer shared PAVs (8,277) among species/accessions may suggest potential impacts of SVs on species/accession-specific traits (relative to *Gm*; **Supplementary Fig. 9a and c**).

The number of PAVs in the Dt genome (range from 61,132 to 67,223) is slightly smaller than that in the At genome (range from 64,875 to 78,695) for all polyploid cotton accessions except *Gh*, suggesting a higher density of PAVs in the much smaller Dt subgenome (**Supplementary Table 14**). Most PAVs were located in intergenic regions (70.53%–76.81%) and were lower in coding regions than in introns (**Supplementary Table 16 and Supplementary Fig. 10**), as expected. 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) (**Supplementary Table 17**), including 1,168 proteins affected in at least four samples simultaneously. Corresponding to 8,277 SVs shared by all accessions relative to *Gm*, a total of 646 protein sequences were affected, exhibiting the highest rate of genic changes by SVs in the phylogenetic tree (0.078); on the terminal branches, species/accession-specific rates ranged from 0.006 to 0.027 (**Supplementary Fig. 9b and c**) The three chromosomes most affected by PAVs within genes were At05 (535 genes), At11 (428), and Dt11 (312) (**Supplementary Fig. 11**).

Of the SVs affecting genes, we found a 450-bp deletion/insertion (coordinates in domesticated *G. hirsutum;* A10:84,877,673-84,878,123) that resulted in a shorter version of *Ghi_A10G09231* in *Gb*, *Gd* and *Gt* (**Fig. 3a**). This gene encodes a phosphopeptide-binding protein[47] that is involved in fiber length, and which exhibits significantly reduced expression in fiber from cotton species missing this gene fragment (**Supplementary Fig. 12**). This deletion (relative to *Gm*) was confirmed to be missing in *Gb* and *Gd*, suggesting that the deletion occurred subsequent to divergence from *Gm*.

We also found a large-scale inversion event (~ 4.48 Mb) on D04 that distinguishes most of the *Gh*-like (i.e., *Ge, Gs,* and *Gh)* from *Gt* and the *Gb*-like clade (**Fig. 3b**). This inversion was further confirmed by mapping Hi-C data of four accessions (*Gb*, *Ge*, *Gs* and *GhP*) to TM-1_WHU, as shown in **Supplementary Fig. 13**. This inversion is phylogenetically inferred to have occurred prior to divergence among the closely-related terminal species in the *Gh*-like clade, and may have contributed to genetic isolation of *Gh*-like species and the *Gb*-like clade.

We asked whether both *Gb* and *Gh* underwent convergent domestication by inspecting similar genomic changes during domestication [48]. Remarkably, an inversion event larger than 986.42 Kb on D01 was observed in both domesticated *Gb* and *Gh* **(Fig. 3c**), suggesting either a remarkable convergence under human selection, or more likely introgression between the two species, which has been common[12].

## A resource for disease resistance and stress response discovery

*Disease resistance*

Plant resistance to biotic stressors, such as pathogens and pests, is usually mediated by disease resistance genes, most of which encode intracellular nucleotide-binding site leucine-rich repeat (NLR)

proteins[49, 50]. These proteins take part in innate immunity by directly or indirectly recognizing pathogen effector proteins[51], and previous studies have associated more NLRs with broader-spectrum disease resistance. The NLR gene content of all seven tetraploid species (8 accessions) was investigated using a disease resistance gene analog prediction pipeline, which identified 3,462 to 4,312 NLR genes in each of the eight cotton genomes (**Supplementary Table 18**). Notably, NLR content was significantly higher in all Dt subgenomes (versus At; *t test*: $p$=0.012), congruent with earlier reports for 5 allopolyploid cotton species[18] (**Supplementary Fig. 14**). The NLR gene clusters were scattered across the almost all chromosomes, with dense clusters appearing in A04, A11, and D11 of *GhP* and *Ge* (**Supplementary Fig. 15**). Oligonucleotide probes designed for *Ge* A04 and A11 R gene clusters (see methods) confirmed the presence of these clusters in other tetraploid genomes **(Supplementary Fig. 16)**. Those several distinct cotton R genes clusters on chromosomes, with high similarity between sequences, suggested some R gene cluster structure occurred prior to genomic differentiation of tetraploid cotton.

*Drought and salt tolerance*

Wild species of tetraploid cotton naturally occur in habitats that periodically are subjected to drought and/or salt-stress[7, 52]. Thus, they likely harbor potentially useful genes for adaptation to these challenges, and, to this extent, the genomes presented here provide resources for gene discovery[33, 53]. In *GhP*, 459 specific gene families were found to be enriched in “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 (ath00061, $p$= 6.12e-05)” (**Supplementary Figs. 17 and 18**), which may play a functional role in biotic and abiotic stress tolerance. Among these genes is *GhECI3* (*GhirPD0101G028900*), which encodes a homologue of enoyl-CoA delta isomerase 3 that is involved in salt and drought stress response in *Arabidopsis thaliana*[54] **(Fig. 4b**). By exploring the *GhP* transcriptome under salt and drought stress (**Fig. 4a, Supplementary Table 19**), we confirmed that expression of *GhECI3* was significantly decreased in the early stages of either cold or salt stress (**Fig. 4d-h**). We also found that differentially expressed genes (DEGs) in response to these stresses were enriched in carbohydrate metabolic processes (**Fig. 4b-c**). Interestingly, the number of DEGs responding to drought were fewer than those responding to salt, which may indicate a more streamlined response to drought stress and/or more pleiotropic effects accompanying the response to drought stress (**Fig. 4a**). These results exemplify the potential of the new assemblies presented here for gene discovery.

We speculate that drought and salt stress responsive expression changes may be shaped by selection. To test this hypothesis, we inferred 369 positively selected genes (PSGs) in *GhP*, including 188 and 181 in the At and Dt subgenomes, respectively. Using the combined RNA-seq data sets (control and stress treated) from above, we directly compared gene expression levels between PSGs and other single-copy genes (SCGs). Interestingly, we found significantly higher PSG than SCG expression in the Dt subgenome of the primitively domesticated *GhP* (**Supplementary Fig. 19a**), whereas SCG exhibited higher expression than PSG in both At and Dt subgenomes of *Gb* (**Supplementary Fig. 19b**). These results indicate lineage-specific asymmetrical expression evolution in cotton subgenomes.

We note that one of the PSGs, *GhdadD* (*GhirPA0101T216900*), may be related to abiotic stress response[55] as it encodes a Phospholipase A, which catalyzes the initial step of jasmonic acid biosynthesis. Expression of this gene was down-regulated under salt and drought stress (**Supplementary Fig. 20a**), and both salt and drought tolerance were decreased in *GhdadD* knock-down seedlings (**Supplementary Fig. 20b, c and d**).

## Homoeologous exchange (HE) reveals asymmetrical evolution

Homoeologous exchanges are increasingly recognized as being widespread following polyploidy in plants, often changing relative homoeolog dosages and creating phenotypically variable progeny[25, 26]. To explore this in *Gossypium* polyploids, we used the $D_5$ and $A_1$ diploid genomes as models of the Dt and At subgenome donors, respectively[37]. We define unexpectedly close gene pairs between the subgenomes of the tetraploid and their *non*-donor diploid genome as AtD (i.e., orthologs between At subgenome and $D_5$) and DtA (i.e., orthologs between Dt subgenome and $A_1$), representing discordant orthologous relationships that are likely caused by HEs. Reciprocal protein alignments were used to detect the best protein sequence homology in eight tetraploids tracing to the same allopolyploidization event, with $D_5$ and $A_1$ used as models of the diploid genomes. Expected orthologous relationships, i.e., AtA and DtD, accounted for 81-83% of all inferred relationships (**Fig. 5a**), far more than discordant relationships, AtD and DtA. In addition, we observed significantly higher numbers of AtD than DtA in all eight tetraploid cottons (**Supplementary Table 20 and Supplementary Fig. 21**). These results indicate that asymmetrical HEs may have rendered genes of D-genome origin to be A-like more often than the HEs in the other direction.

We next characterized the potential HE events for each tetraploid cotton genome based on the enrichment of discordant orthologous pairs (Supplementary Methods). In brief, a HE event was diagnosed when a significantly high number of AtD and DtA genes was detected in windows of more than 50 Kb, and the coverage depth of above-mentioned windows by tetraploid illumina sequences were 1.5 times that of the 10kb windows on both sides. This approach allowed us to identify a total of 51 HE events, including 10 from At to Dt (DtA) and 41 from Dt to At (AtD), with an average summed length of 8.5 Mb in each genome and an average length of 107.94 Kb (**Supplementary Tables 21-22 and Supplementary Fig. 22**). We found that 11 HE events (all from Dt to At) were shared among the eight tetraploid cotton genomes, and 6 HE events were lineage-specific (2 specific to *Gs*, and 1 each specific to *Gb*, *Gt*, *Gd*, and *Gh*; **Supplementary Fig. 23**). We thus concluded that after polyploid formation, non-reciprocal HEs occurred between At and Dt favoring the direction of Dt to At conversion. Further, we found all HE events overlapped with TE regions (**Supplementary Tables 22-23)**, suggesting that TEs are the driving force for the occurrence of HE events in tetraploid cottons.

Notably, we identified 152 genes with characterization of leucine rich repeats (LRRs) significantly enriched in HE regions (*fisher test*, $p$=3.9e-113). This finding implies that natural selection shaped genome after the formation of tetraploid. Interestingly, we found a 180 Kb HE region result from homologous exchange of D01 to A01 in *Gh*, with obvious low *Ks* value to D-ancestral and high coverage

by TM-1 Illumina reads. We further found that this HE event is present in all eight tetraploid cotton genomes (**Supplementary Fig. 24**), and that it contains an RGA (resistance gene analog) gene encoding a putative receptor-like protein kinase, possibly reflecting differential selection for resistance to diseases.

# Discussion

we present three *de novo* assemblies of cotton genomes, *Gs*, *Gm*, and *GhP*, to complete the phylogenic representation of all seven tetraploid *Gossypium* species and also provide an improved foundation for understanding the domestication of upland cotton, *Gh*. Our research confirmed a monophyletic origin of tetraploid cottons and reiterated asymmetric evolution of the co-resident At and Dt subgenomes. Analyses of genome-wide structural variation revealed a large inversion in chromosome D01 common to domesticated forms of *Gh* and *Gb*, representing either convergent domestication or introgression between upland and sea island cottons. By examining transcriptional responses to abiotic and biotic stresses, we demonstrated the use of these tetraploid genomes to further functional discoveries in cotton. HE analysis revealed that directional exchange from the Dt to the At subgenomes occurred more frequently than the reverse in all eight tetraploid cotton genomes. A significant enrichment of LRR genes in HE regions is suggestive of potential functional consequences of differential genome evolution during speciation following allopolyploid formation. Our study provides a valuable resource for polyploid genome evolution and for understanding crop domestication, as well as for heterosis and functional genomics to facilitate cotton breeding.

# Methods

### Plant materials and growth conditions

Plant materials used in this study for DNA sequencing were leaves, collected from following three cotton species: *Gossypium ekmanianum* accession *AD602*, *Gossypium stephensii* accession number *AD701*, and *Gossypium hirsutum* race *Punctatum* accession number *Punctatum 25* (TX-1000). These plants are all perennially maintained at the National Wild Cotton Nursery in Sanya, China, which is supervised by 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 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 Ⅱ 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[56]. Briefly, the leaves of 15-day-old seedlings were fixed in 1% formaldehyde solution. The nuclei/chromatin was extracted from the fixed tissue and digested with *DpnII*. The overhangs resulted from *DpnII* digestion were filled in using biotin-14-dCTP (Invitrogen) and the Klenow enzyme (NEB). After dilution and relegation chromatin with T4 DNA ligase (NEB), genomic DNA was extracted and sheared to a size of 300 to 500 bp with Bioruptor (Diagenode). The biotin-labeled DNA fragments were enriched using streptavidin beads (Invitrogen) and subject to library preparation according previous report[57]. Illumina sequencers (Illumina HiSeq X Ten platform) carried out the sequencing of the Hi-C libraries. HiC-Pro (v.2.10.0) was used to evaluate Hi-C data quality[57]. Samples from leaves, stems, and stem apices of mature *GhP*, *Gs,* and *Ge* plants were collected for extracting RNA. Then we constructed RNA-seq libraries using the protocol of NEBNext® Ultra™ RNA Library Prep Kit for Illumina® (NEB, USA). RNA-seq libraries were also sequenced using Illumina X Ten platform.

### Contig assembly using PacBio reads

De novo genome assembly was performed mainly using the PacBio SMART long reads with FALCON[58] (https://github.com/PacificBiosciences/FALCON/, falcon-kit==1.8.1). Briefly, we first selected the longest 50 X 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[59] (https://www.pacb.com/support/software-downloads) by aligning total SMRT reads. Lastly, Pilon (V1.18) [60] were used to perform the 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: (a) Reads with ≥ 10% unidentified nucleotides (N); (b) Reads with > 10 nt aligned to the adapter, allowing ≤ 10% mismatches; (c) Reads with > 50% bases having phred quality < 5. The filtered Hi-C reads were aligned against the contig assemblies with BWA (version 0.7.8). 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 HiC read-pairs linking each pair of scaffolds was tabulated. LACHESIS[61] (https://github.com/shendurelab/LACHESIS) used hierarchical agglomerative clustering to twenty-six groups. Juicebox v1.22 (https://github.com/aidenlab/Juicebox) were finally used to order the scaffolds in each group.

### Genome assembly quality assessment

The genome assembly was evaluated by mapping the high-quality reads from 350bp 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 [62] (Benchmarking Universal Single-Copy Orthologs, version 3.0.2) was used to evaluate the assembly completeness of three cotton genome with 1,440 embryophyte genes from the 'Embryophyta_odb9' database. LAI (LTR

Assembly Index) was used to evaluate assembly continuity and completeness by full-length long terminal repeats retrotransposons (LTR-RTs). LTRharvest (V1.5.3)[63] (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)[64] (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 whole three cottons genome assembly, respectively. LTR_retriever (V2.9.0)[65] was then used to combine and refine all the candidates to get the final completeness LTR-RTs. The 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 (version 1.0.8), RepeatScout (version 1.0.5), and LTR_FINDER (version 1.07) 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 (version 3.3.0) to predict TEs. RepeatProteinMask (version 3.3.0, http://www.repeatmasker.org/RepeatMasker) which makes homology-based predictions, was performed to detect TEs in these three cottons genome by comparing them against to the TE protein database. Tandem repeats were detected in the genome using Tandem Repeats Finder (TRF, version 4.07b).

## Protein-coding gene annotation

A combination of *de novo*, homology-based, and RNA-seq based predictions were employed to annotate the protein-coding genes in the three cottons genomes. Five *ab initio* gene prediction programs were used to predict genes, including Augustus[66] (version 3.0.2), Genescan[67] (version 1.0), Geneid[68] (version 1.4), GlimmerHMM[69] (version 3.0.2) and SNAP[70] (version 2013-02-16). Protein sequences from six dicot species (i.e., *Arabidopsis thaliana[71]*, *Theobroma cacao[72]*, *Populus trichocarpa[73]*, *Gossypium hirsutum[17]*, *Gossypium arboreum*[17], and *Gossypium raimondii*[35]) were downloaded from cottongen[74], Ensembl[75] and NCBI[76] and aligned against to the genome using WUblast (version 2.0) [77]. Genewise[78] (version 2.2.0) was employed to predict gene models based on the sequence alignment results. For RNA-seq based predictions, more than four tissue (root, stem, leaf, flower and so on) RNA-seq data were aligned to the three cottons genome using Tophat[79] (version 2.0.13) to identify exons region and splice positions. The alignment results were then used as input for cufflinks[80] (version 2.1.1) to assemble transcripts to the gene models. In addition, the RNA-seq data was assembled by Trinity (version 2.1.1), creating several pseudo-ESTs, which were mapped to the three cottons assembly genome by BLAT (V3.2.3)[81] and used to predict gene models via PASA (r20140417)[82]. A weighted and non-redundant gene set was generated by EVidenceModeler [83] (EVM, version 1.1.1) which merged all genes models predicted by the above three approaches. Combining with transcript assembly, PASA adjusted the gene models generated by EVM.

## Functional annotation

The predicted protein sequences were assigned functions by searching six protein/function databases: NR, InterPro, GO, KEGG, Swiss-Prot, and TrEMBL. We used InterproScan46 (v20180213)[84] 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. R package (Clusterprofiler47[85]) R package ()were used to do the GO term and KEGG enrichment analysis.

## Orthology and Pan-genome analysis

Protein sequences of annotated genes from eight diploid cottons (*Gki*[86], *Gau*[33], *Glo*[32], *Ghe*[17], *Gar*[31] , *Glo*[32], *Gra*[35]) and eight allotetraploid cottons (*Gh*[18], *Gb*[18], *Gt*[18], *Gm*[18], *Gd*[18], *Ge*, *Gs*, *GhP*) were merged. The longest proteins of each gene were used to align to themselves by BLASTP with e-value cutoff of 1e-5. OrthoFinder (v2.2.7)[87] was applied to detect orthogroups for all homologous genes across the merged protein sequences with default parameters. Single copy genes among those cotton genomes were first aligned by MUSCLE (v3.8.31)[88] in each gene cluster and concatenated into a super-alignment. 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". *Ka* 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)[89] software. Divergence times were estimated using the mathematical formula T = Ks/2r (substitution rate r = $2.6 \times 10^{-9}$)[90].

Putative positively selected genes were detected using the branch-site model in PAML (V4.7)[91]. Genome synteny blocks containing at least four genes was detected using mcscan (https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version) with parameter: –cscore=.90, –iter=1. Gene families for the eight tetraploid cottons (*Gh*, *Gb*, *Gt*, *Gm*, *Gd*, *Ge*, *Gs*, *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 was defined as species-special gene families. The gene families that were presenting in one to seven samples were defined as dispensable gene families.

## Resistance gene analogs (RGAs) identification and evolution

To predict RGAs in cotton tetraploids genome, RGAdb from RGAugury (https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-3197-x)[92] software was downloaded. Protein sequences of all annotated genes of cottons were aligned to the RGAdb using BLASTP with an e-value cutoff of 1e-05. Seven RGAs-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.

## RNA-seq for samples under salt and drought stress

All the samples using for RNA-Seq were collected from National Wild Cotton Nursery in Sanya, China. Four-week-old seedlings of *GhP* and *Gh* (TM-1) were exposed to both salt (300mM NaCl) and PEG

(200g/L PEG). Leaf samples were collected post treatment at 0 hours, 12 hours, and 24 hours. All fresh tissues were frozen in liquid nitrogen and stored at -80°C before processing. Total RNA for each sample was 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 Inc. San Diego, CA, USA) and sequenced on an Illumina NovaSeq platform using pair-end short reads (150bp) sequencing strategy.

RNA-seq data were mapped to the corresponding genome using Tophat2 (v2.0.8)[79], and HTSeq[93] v0.6.1 was used to count the number of reads mapped to each gene. FPKM 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 performed using the DESeq R package[94] (1.18.0). Genes with an adjusted P-value <0.05 were considered differentially expressed.

## Genomic SVs detection

We aligned seven allotetraploid cotton genomes to the *Gm* (AD4_JGI) reference genome and then combined three methods to identify SVs including smartie-SV[43] (https://github.com/zeeev/smartie-sv), SyRI[45] (https://github.com/schneebergerlab/syri) and SVMU[44] (https://github.com/mahulchak/svmu). Specifically, the pipeline of Smartie-SV was preformed based on the BLASR (V5.3.2) alignment with default parameters; we extracted alignment pairs from any pair of genomes based on nucmer (V 3.23)[95] (–mum –maxgap=500 –mincluster=1000) to serve as input for the packages of SyRI and SVMU with default parameter. Then, we aligned Illumina reads of seven tetraploid cotton genomes to the TM-1_WHU reference genome to identify SVs using Breakdancer (version 1.3.6). On the basis of the above pipeline, we obtained four raw SV sets. For insertions and deletions, we merged four raw set using package Jasmine[23] (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. For inversions, we also only considered candidates supported by at least two methods by applying software bedtools[96]. Annotation of genomic SVs was performed using the package ANNOVAR[97] (Version: 2019Oct24). Based on the genome annotation, 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.

## Identification of HE events

For each allotetraploid cotton genome, we performed bidirectional alignment using only protein sequences with more than 50 amino acids in an all-against-all BLASTP (E-value: 1e-5). We determined gene similarities between the subgenome of tetraploid cotton genome and its donor diploid genomes, thereby build a similarity graph of protein-coding genes among tetraploid cotton genomes and their diploid models *G. herbaceum* (A1) and *G. raimondii* (D5). Subsequently, in order to identify

orthologous gene pairs, we extracted alignment pairs from any pair of genomes and restricted a maximum of five hits per protein sequence to serve as input for the MCScanX algorithm[98] with the parameters 'MATCH_SCORE: 50, MATCH_SIZE: 5, GAP_SCORE:−3, E_VALUE: 1e-05', which was used to detect high-confidence collinear blocks of coding genes and to identify orthologous gene pairs. Next, we defined orthologous gene pairs between the subgenome of the tetraploids and their non-donor diploid genome as 'AHGP' (i.e., AtD, that is orthologs between At subgenome and $D_5$, and DtA, that is orthologs between Dt subgenome and $A_1$), and defined orthologous gene pairs between the subgenome of the tetraploids and their donor diploid genome as 'HHGP' (i.e., AtA and DtD, genes in the two subgenomes closer to their presumed diploid donor homologs). We counted the number of AHGP and HHGP with 50 Kb non-overlapping windows along the genomes. We considered a region to exhibit HE when the number of AHGP genes was greater than HHGP genes, and we merged adjacent HE windows. Subsequently, we mapped Illumina pair-end reads of the tetraploid cottons to a combined genome reference composed of A1 and D5 using BWA. We selected only uniquely mapping reads to determine read depth for the A1 and D5 genomes. This ensures a tetraploid cotton read can map only once onto either the At or Dt genome. When the coverage depth of putative HE regions was 1.5 times that of the 10kb windows on both sides, we considered this region to be the result of HE.

## Virus induced gene silencing (VIGS)

VIGS of the genes *GhECI3* and *GhdadD* 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[99]. Firstly, the VIGS vectors TRV:*GhECI3* and TRV:*GhdadD* were constructed by recombine approximately 300-bp fragments of *GhECI3* and *GhdadD* into pTRV-RNA2 vector, and introduced 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 previous protocol[33]. The transformed cotton seedlings were grown under greenhouse conditions, which were 25℃ and 8h dark/16h day cycle. After 20 days post *Agrobacterium* inoculation, the VIGS-plants and non-VIGS plants were exposed to salt (300mM NaCl) and drought (17% PEG6000) treatment for 3 days. Finally, we collected the leaves of TRV:*GhECI3*, TRV:*GhdadD*, TRV:00 and non-VIGS seedlings for morphological and physiological analysis.

## Oligo probes

Six probes for RGAs were designed based on the *Gossypium ekmanianum* genome sequence. These oligo probes were synthesized by Ningbo Kangbei Biochem, Inc. (Ningbo, China), which attached a 6-carboxyfluorescein (6-FAM) or 6-carboxytetramethylrho-damine (TAMRA) to the 5' end. The primer sequence information is shown in **supplementary table 24**.

The oligo probes were designed according to previous methods[100]. In short, the RGA sequences enriched in the chromosomes A04, A11 and D11 of *G. ekmanianum* were analyzed using the Tandem Repeats Finder (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, 20–60 and > 60). At the same time, the tandem repeats were physical mapped onto the genome sequence using a web server B2DSC (http://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 (version 22.0, SPSS, Chicago, IL).

### FISH analysis of RGA-derived Oligo probes

Root tips of five cotton species: *Gossypium hirsutum* (cultivar: TM-1), *Gossypium barbadense* (cultivar: 3-79), *Gossypium tomentosum* (accession number in ICR-CAAS: AD3-LZ), *Gossypium mustelinum* (accession number in ICR-CAAS: AD4-LZ), and *Gossypium darwinii* accession (accession number in ICR-CAAS: AD5-07), were harvested from circa 6-day-old incubator-grown seedlings. Root tips were pretreated using 25 ppm 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 previously reported methods[101]. The protocol of ND-FISH using synthesized probes was described by Tang et al. (2018)[102]. In short, 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℃ for 1-3 h. After hybridization, the slides were placed in 2×SSC solution until the plastic film cover fell off naturally. Slides were dried in the dark. Chromosomes were counter-stained with 4’-6-diamidino-2-phenylindole (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.

# Declarations

### Data availability

All the raw sequencing data are available in the National Center for Biotechnology Information’s (NCBI) BioProject database (accession no. PRJNA739494). The genome assemblies and annotation files are available at NCBI genome database (accession no. SAMN19790799- SAMN19790801).

### Acknowledgements

We thank the National Natural Science Foundation of China (31621005, 32072023, 31471548), the National Key R&D Program of China (2021YFE0101200), PSF/CRP/18thProtocol (07) and Agricultural Science and Technology Innovation Program of Chinese Academy of Agricultural Sciences for financially supporting.

### Author information

These authors contributed equally: Renhai Peng, Yanchao Xu, Zhen Liu, Liyang Chen, Zhongli Zhou, Xiaoyan Cai

## Affiliations

### Research Base, Anyang Institute of Technology, State Key Laboratory of Cotton Biology, Anyang, 455000, China

Renhai Peng, Zhen Liu, Yangyang Wei, Yuling Liu, Pengtao Li, Tao Wang & Quanwei Lu

### State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences, Anyang, Henan, China

Yanchao Xu, Zhongli Zhou, Xiaoyan Cai, Kunbo Wang, Heng Wang, Guanjing Hu, Yuqing Hou, Yuhong Wang & Fang Liu

### Anyang Institute of Technology, Anyang, 455000, China

Renhai Peng, Zhen Liu, Yangyang Wei, Yuling Liu, Pengtao Li, Tao Wang & Quanwei Lu

### Novogene Bioinformatics Institute, Beijing, 100015, China

Liyang Chen & Shilin Tian

### Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50011, USA

Corrinne E. Grover & Jonathan F. Wendel

### Collaborative Innovation Center of Modern Biological Breeding, School of Life Science and Technology, Henan Institute of Science and Technology, Xinxiang, China

Yuanyuan Wang & Qinlian Wang

### School of Agricultural Sciences, Zhengzhou University, Zhengzhou, Henan, 450001, Peoples republic of China

Fang Liu

## Contributions

F.L., R.P., and S.T., conceived and designed the project. Z.Z., X.C. and Y.X. collected materials. L.C. and S.T. generated the PacBio and Illumina sequencing data. L.C., S.T. Z.L. and Y.X. analyzed the data. K.W., Y.W., Y.L., H.W., Y.H., Y.W., P.L., Y.X., T.W., Q.L., and Y.W. performed experiments. F.L., J.W., S.T., K.W., Z.Z., H.G., C.G., Y.X., and Q.W. contributed to project discussion. Y.X. and S.T. wrote the manuscript draft. F.L., J.W., C.G. and G.H. revised it.

### Ethics declarations

### Competing interests

The authors declare no competing interests.

# References


1. Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, et al: **Ancestral polyploidy in seed plants and angiosperms.** *Nature* 2011, **473:**97-100.

2. Comai L: **The advantages and disadvantages of being polyploid.** *Nat Rev Genet* 2005, **6:**836-846.

3. Otto SP: **The evolutionary consequences of polyploidy.** *Cell* 2007, **131:**452-462.

4. Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, Depamphilis CW, Wall PK, Soltis PS: **Polyploidy and angiosperm diversification.** *Am J Bot* 2009, **96:**336-348.

5. Van de Peer Y, Mizrachi E, Marchal K: **The evolutionary significance of polyploidy.** *Nat Rev Genet* 2017, **18:**411-424.

6. Wendel JF, Flagel LE, Adams KL: **Jeans, Genes, and Genomes: Cotton as a Model for Studying Polyploidy.** In *Polyploidy and Genome Evolution.* Edited by Soltis PS, Soltis DE. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012: 181-207

7. Wendel J, Grover C: **Taxonomy and Evolution of the Cotton Genus, Gossypium.** In. Edited by Fang DD, Percy RG; 2015

8. Gallagher JP, Grover CE, Rex K, Moran M, Wendel JF: **A New Species of Cotton from Wake Atoll, Gossypium stephensii (Malvaceae).** *Systematic Botany* 2017, **42:**115-123.

9. Grover CE, Zhu X, Grupp KK, Jareczek JJ, Gallagher JP, Szadkowski E, Seijo JG, Wendel JF: **Molecular confirmation of species status for the allopolyploid cotton species, Gossypium ekmanianum Wittmack.** *Genetic Resources and Crop Evolution* 2015, **62:**103-114.

10. Soltis DE, Visger CJ, Marchant DB, Soltis PS: **Polyploidy: Pitfalls and paths to a paradigm.** *American Journal of Botany* 2016, **103:**1146-1166.

11. Chen ZJ: **Molecular mechanisms of polyploidy and hybrid vigor.** *Trends in Plant Science* 2010, **15:**57-71.

12. Yuan D, Grover CE, Hu G, Pan M, Miller ER, Conover JL, Hunt SP, Udall JA, Wendel JF: **Parallel and Intertwining Threads of Domestication in Allopolyploid Cotton.** *Advanced Science* 2021, **8:**2003634.

13. Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J, et al: **Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres.** *Nature* 2012, **492:**423-427.

14. Kunbo W, Jonathan W: **Designations for individual genomes and chromosomes in Gossypium.** *Journal of Cotton Research* 2018, **1:**3.

15. Wang M, Tu L, Yuan D, Zhu, Shen C, Li J, Liu F, Pei L, Wang P, Zhao G, et al: **Reference genome sequences of two cultivated allotetraploid cottons, Gossypium hirsutum and Gossypium barbadense.** *Nat Genet* 2019, **51:**224-229.

16. Yang Z, Ge X, Yang Z, Qin W, Sun G, Wang Z, Li Z, Liu J, Wu J, Wang Y, et al: **Extensive intraspecific gene order and gene structural variations in upland cotton cultivars.** *Nat Commun* 2019, **10:**2989.

17. Huang G, Wu Z, Percy RG, Bai M, Li Y, Frelichowski JE, Hu J, Wang K, Yu JZ, Zhu Y: **Genome sequence of Gossypium herbaceum and genome updates of Gossypium arboreum and Gossypium hirsutum provide insights into cotton A-genome evolution.** *Nat Genet* 2020, **52:**516–524.

18. Chen ZJ, Sreedasyam A, Ando A, Song Q, De Santiago LM, Hulse-Kemp AM, Ding M, Ye W, Kirkbride RC, Jenkins J, et al: **Genomic diversifications of five Gossypium allopolyploid species and their impact on cotton improvement.** *Nat Genet* 2020.

19. Hu Y, Chen J, Fang L, Zhang Z, Ma W, Niu Y, Ju L, Deng J, Zhao T, Lian J, et al: **Gossypium barbadense and Gossypium hirsutum genomes provide insights into the origin and evolution of allotetraploid cotton.** *Nat Genet* 2019, **51:**739-748.

20. Kou Y, Liao Y, Toivainen T, Lv Y, Tian X, Emerson JJ, Gaut BS, Zhou Y: **Evolutionary Genomics of Structural Variation in Asian Rice (Oryza sativa) Domestication.** *Mol Biol Evol* 2020, **37:**3507-3524.

21. Nieto Feliner G, Casacuberta J, Wendel JF: **Genomics of Evolutionary Novelty in Hybrids and Polyploids.** *Front Genet* 2020, **11:**792.

22. Zhou Y, Massonnet M, Sanjak JS, Cantu D, Gaut BS: **Evolutionary genomics of grape (Vitis vinifera ssp. vinifera) domestication.** *Proc Natl Acad Sci U S A* 2017, **114:**11715-11720.

23. Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D, et al: **Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato.** *Cell* 2020, **182:**145-161 e123.

24. Xiong Z, Gaeta RT, Pires JC: **Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid Brassica napus.** *Proc Natl Acad Sci U S A* 2011, **108:**7908-7913.

25. Mason AS, Wendel JF: **Homoeologous Exchanges, Segmental Allopolyploidy, and Polyploid Genome Evolution.** *Front Genet* 2020, **11:**1014.

26. Wu Y, Lin F, Zhou Y, Wang J, Sun S, Wang B, Zhang Z, Li G, Lin X, Wang X, et al: **Genomic mosaicism due to homoeologous exchange generates extensive phenotypic diversity in nascent allopolyploids.** *National Science Review* 2020, **8:**nwaa277.

27. Page JT, Liechty ZS, Alexander RH, Clemons K, Hulse-Kemp AM, Ashrafi H, Van Deynze A, Stelly DM, Udall JA: **DNA Sequence Evolution and Rare Homoeologous Conversion in Tetraploid Cotton.** *PLoS Genet* 2016, **12:**e1006012.

28. Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, Rausch T, Stutz AM, Stedman W, Anantharaman T, Hastie A, et al: **Assembly and diploid architecture of an individual human genome via single-molecule technologies.** *Nat Methods* 2015, **12:**780-786.

29. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics* 2015, **31:**3210-3212.

30. Ou S, Chen J, Jiang N: **Assessing genome assembly quality using the LTR Assembly Index (LAI).** *Nucleic Acids Res* 2018, **46:**e126.

31. Du X, Huang G, He S, Yang Z, Sun G, Ma X, Li N, Zhang X, Sun J, Liu M, et al: **Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits.** *Nat Genet* 2018, **50:**796-802.

32. Grover CE, Pan M, Yuan D, Arick MA, Hu G, Brase L, Stelly DM, Lu Z, Schmitz RJ, Peterson DG, et al: **The Gossypium longicalyx Genome as a Resource for Cotton Breeding and Evolution.** *G3 (Bethesda)* 2020, **10:**1457-1467.

33. Cai YF, Cai XY, Wang QL, Wang P, Zhang Y, Cai CW, Xu YC, Wang KB, Zhou ZL, Wang CX, et al: **Genome sequencing of the Australian wild diploid species Gossypium australe highlights disease resistance and delayed gland morphogenesis.** *Plant Biotechnology Journal* 2020, **18:**814-828.

34. Grover CE, Arick MA, 2nd, Thrash A, Conover JL, Sanders WS, Peterson DG, Frelichowski JE, Scheffler JA, Scheffler BE, Wendel JF: **Insights into the Evolution of the New World Diploid Cottons (Gossypium, Subgenus Houzingenia) Based on Genome Sequencing.** *Genome Biol Evol* 2019, **11:**53-71.

35. Udall JA, Long E, Hanson C, Yuan D, Ramaraj T, Conover JL, Gong L, Arick MA, Grover CE, Peterson DG, Wendel JF: **De Novo Genome Sequence Assemblies of Gossypium raimondii and Gossypium turneri.** *G3 (Bethesda)* 2019, **9:**3079-3085.

36. Seelanan T, Schnabel A, Wendel JF: **Congruence and consensus in the cotton tribe (Malvaceae).** *Systematic Botany* 1997, **22:**259-290.

37. Wendel JF: **New World tetraploid cottons contain Old World cytoplasm.** *Proc Natl Acad Sci U S A* 1989, **86:**4132-4136.

38. Stewart JM, Oosterhuis D, Heitholt JJ, Mauney JR: *Physiology of cotton.* Springer Science & Business Media; 2009.

39. Brubaker, Curt, L., Wendel, Jonathan, F.: **Reevaluating the origin of domesticated cotton (Gossypium hirsutum; Malvaceae) using nuclear.** *American Journal of Botany* 1994, **81:**1309-1326.

40. Wendel JF, Percy RG: **Allozyme diversity and introgression in the Galapagos Islands endemic Gossypium darwinii and its relationship to continental G. barbadense.** *Biochemical Systematics and Ecology* 1990, **18:**517-528.

41. Golicz AA, Batley J, Edwards D: **Towards plant pangenomics.** *Plant Biotechnol J* 2016, **14:**1099-1105.

42. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH, et al: **An integrated map of structural variation in 2,504 human genomes.** *Nature* 2015, **526:**75-81.

43. Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJP, Dougherty ML, et al: **High-resolution comparative analysis of great ape genomes.** *Science* 2018, **360:**eaar6343.

44. Chakraborty M, Emerson JJ, Macdonald SJ, Long AD: **Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits.** *Nat Commun* 2019, **10:**4872.

45. Goel M, Sun H, Jiao WB, Schneeberger K: **SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies.** *Genome Biol* 2019, **20:**277.

46. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al: **BreakDancer: an algorithm for high-resolution mapping of genomic structural variation.** *Nat Methods* 2009, **6:**677-682.

47. Zhou Y, Zhang ZT, Li M, Wei XZ, Li XJ, Li BY, Li XB: **Cotton (Gossypium hirsutum) 14-3-3 proteins participate in regulation of fibre initiation and elongation by modulating brassinosteroid signalling.** *Plant Biotechnol J* 2015, **13:**269-280.

48. Wang M, Li W, Fang C, Xu F, Liu Y, Wang Z, Yang R, Zhang M, Liu S, Lu S, et al: **Parallel selection on a dormancy gene during domestication of crops from multiple families.** *Nat Genet* 2018, **50:**1435-1441.

49. Jung HW, Panigrahi GK, Jung GY, Lee YJ, Shin KH, Sahoo A, Choi ES, Lee E, Man Kim K, Yang SH, 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* 2020, **32:**1081-1101.

50. Andersen EJ, Ali S, Byamukama E, Yen Y, Nepal MP: **Disease Resistance Mechanisms in Plants.** *Genes (Basel)* 2018, **9:**339.

51. Wang L, Zhao L, Zhang X, Zhang Q, Jia Y, Wang G, Li S, Tian D, Li WH, Yang S: **Large-scale identification and functional analysis of NLR genes in blast resistance in the Tetep rice genome sequence.** *Proc Natl Acad Sci U S A* 2019, **116:**18479-18487.

52. Fryxell PA: *The Natural History of the Cotton Tribe (Malvaceae, Tribe Gossypieae).* College Station, Texas: Texas A & M University Press.; 1979.

53. Zeng L, Tu XL, Dai H, Han FM, Lu BS, Wang MS, Nanaei HA, Tajabadipour A, Mansouri M, Li XL, et al: **Whole genomes and transcriptomes reveal adaptation and domestication of pistachio.** *Genome Biol* 2019, **20:**79.

54. Simon, Goepfert, Charles, Vidoudez, Christian, Tellgren-Roth, Syndie, Delessert, J., Kalervo: **Peroxisomal Δ3,Δ2-enoyl CoA isomerases and evolution of cytosolic paralogues in embryophytes.** *Plant Journal* 2008, **5:**728-742.

55. Ishiguro S, Kawai-Oda A, Ueda J, Nishida I, Okada K: **The DEFECTIVE IN ANTHER DEHISCIENCE gene encodes a novel phospholipase A1 catalyzing the initial step of jasmonic acid biosynthesis, which synchronizes pollen maturation, anther dehiscence, and flower opening in Arabidopsis.** *Plant Cell* 2001, **13:**2191-2210.

56. van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker J, Lander ES: **Hi-C: a method to study the three-dimensional architecture of genomes.** *Journal of Visualized Experiments* 2010**:**1869.

57. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, Barillot E: **HiC-Pro: an optimized and flexible pipeline for Hi-C data processing.** *Genome Biol* 2015, **16:**259.

58. Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, et al: **Phased diploid genome assembly with single-molecule real-time sequencing.** *Nat Methods* 2016, **13:**1050-1054.

59. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, et al: **Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data.** *Nat Methods* 2013, **10:**563-569.

60. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM: **Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement.** *PLoS One* 2014, **9:**e112963.

61. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J: **Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions.** *Nat Biotechnol* 2013, **31:**1119-1125.

62. Seppey M, Manni M, Zdobnov EM: **BUSCO: Assessing Genome Assembly and Annotation Completeness.** *Methods Mol Biol* 2019, **1962:**227-245.

63. Ellinghaus D, Kurtz S, Willhoeft U: **LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons.** *BMC Bioinformatics* 2008, **9:**18.

64. Xu Z, Wang H: **LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons.** *Nucleic Acids Res* 2007, **35:**W265-W268.

65. Ou S, Jiang N: **LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons.** *Plant Physiol* 2018, **176:**1410-1422.

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

67. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268:**78-94.

68. Guigo R: **Assembling genes from predicted exons in linear time with dynamic programming.** *J Comput Biol* 1998, **5:**681-702.

69. Majoros WH, Pertea M, Salzberg SL: **TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders.** *Bioinformatics* 2004, **20:**2878-2879.

70. Korf I: **Gene finding in novel genomes.** *BMC Bioinformatics* 2004, **5:**59.

71. Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR: **High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell.** *Nat Commun* 2018, **9:**541.

72. Argout X, Martin G, Droc G, Fouet O, Labadie K, Rivals E, Aury JM, Lanaud C: **The cacao Criollo genome v2.0: an improved version of the genome for genetic and functional genomic studies.** *BMC Genomics* 2017, **18:**730.

73. Hofmeister BT, Denkena J, Colome-Tatche M, Shahryary Y, Hazarika R, Grimwood J, Mamidi S, Jenkins J, Grabowski PP, Sreedasyam A, et al: **A genome assembly and the somatic genetic and epigenetic mutation rate in a wild long-lived perennial Populus trichocarpa.** *Genome Biol* 2020, **21:**259.

74. Yu J, Jung S, Cheng CH, Ficklin SP, Lee T, Zheng P, Jones D, Percy RG, Main D: **CottonGen: a genomics, genetics and breeding database for cotton research.** *Nucleic Acids Res* 2014, **42:**D1229-D1236.

75. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, et al: **Ensembl 2021.** *Nucleic Acids Res* 2021, **49:**D884-D891.

76. Kitts PA, Church DM, Thibaud-Nissen F, Choi J, Hem V, Sapojnikov V, Smith RG, Tatusova T, Xiang C, Zherikov A, et al: **Assembly: a resource for assembled genomes at NCBI.** *Nucleic Acids Res* 2016, **44:**D73-D80.

77. She R, Chu JS, Wang K, Pei J, Chen N: **GenBlastA: enabling BLAST to identify homologous gene sequences.** *Genome Res* 2009, **19:**143-149.

78. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise.** *Genome Res* 2004, **14:**988-995.

79. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.** *Genome Biol* 2013, **14:**R36.

80. Ghosh S, Chan CK: **Analysis of RNA-Seq Data Using TopHat and Cufflinks.** *Methods Mol Biol* 2016, **1374:**339-361.

81. Kent WJ: **BLAT–the BLAST-like alignment tool.** *Genome Res* 2002, **12:**656-664.

82. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, et al: **Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies.** *Nucleic Acids Research* 2003, **31:**5654-5666.

83. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR: **Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments.** *Genome Biology* 2008, **9:**R7.

84. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, et al: **InterPro: the integrative protein signature database.** *Nucleic Acids Res* 2009, **37:**D211-D215.

85. Yu GC, Wang LG, Han YY, He QY: **clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters.** *Omics-a Journal of Integrative Biology* 2012, **16:**284-287.

86. Udall JA, Long E, Ramaraj T, Conover JL, Yuan D, Grover CE, Gong L, Arick MA, 2nd, Masonbrink RE, Peterson DG, Wendel JF: **The Genome Sequence of Gossypioides kirkii Illustrates a Descending Dysploidy in Plants.** *Front Plant Sci* 2019, **10:**1541.

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

88. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32:**1792-1797.

89. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J: **KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies.** *Genomics Proteomics Bioinformatics* 2010, **8:**77-80.

90. Grover CE, Arick MA, 2nd, Conover JL, Thrash A, Hu G, Sanders WS, Hsu CY, Naqvi RZ, Farooq M, Li X, et al: **Comparative Genomics of an Unusual Biogeographic Disjunction in the Cotton Tribe (Gossypieae) Yields Insights into Genome Downsizing.** *Genome Biol Evol* 2017, **9:**3328-3344.

91. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood.** *Mol Biol Evol* 2007, **24:**1586-1591.

92. Li P, Quan X, Jia G, Xiao J, Cloutier S, You FM: **RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants.** *BMC Genomics* 2016, **17:**852.

93. Anders S, Pyl PT, Huber W: **HTSeq-a Python framework to work with high-throughput sequencing data.** *Bioinformatics* 2015, **31:**166-169.

94. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol* 2014, **15:**550.

95. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes.** *Genome Biol* 2004, **5:**R12.

96. Quinlan AR: **BEDTools: The Swiss-Army Tool for Genome Feature Analysis.** *Curr Protoc Bioinformatics* 2014, **47:**11.12.11-11.12.34.

97. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.** *Nucleic Acids Res* 2010, **38:**e164.

98. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, et al: **MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity.** *Nucleic Acids Res* 2012, **40:**e49.

99. Xu Y, Magwanga RO, Yang X, Jin D, Cai X, Hou Y, Wei Y, Zhou Z, Wang K, Liu F: **Genetic regulatory networks for salt-alkali stress in Gossypium hirsutum with differing morphological characteristics.** *BMC Genomics* 2020, **21:**15.

100. Liu XY, Sun S, Wu Y, Zhou Y, Gu SW, Yu HX, Yi CD, Gu MH, Jiang JM, Liu B, et al: **Dual-color oligo-FISH can reveal chromosomal variations and evolution in Oryza species.** *Plant Journal* 2020, **101:**112-121.

101. Liu YL, Wang XY, Wei YY, Liu Z, Lu QW, Liu F, Zhang T, Peng RH: **Chromosome Painting Based on Bulked Oligonucleotides in Cotton.** *Frontiers in Plant Science* 2020, **11:**802.

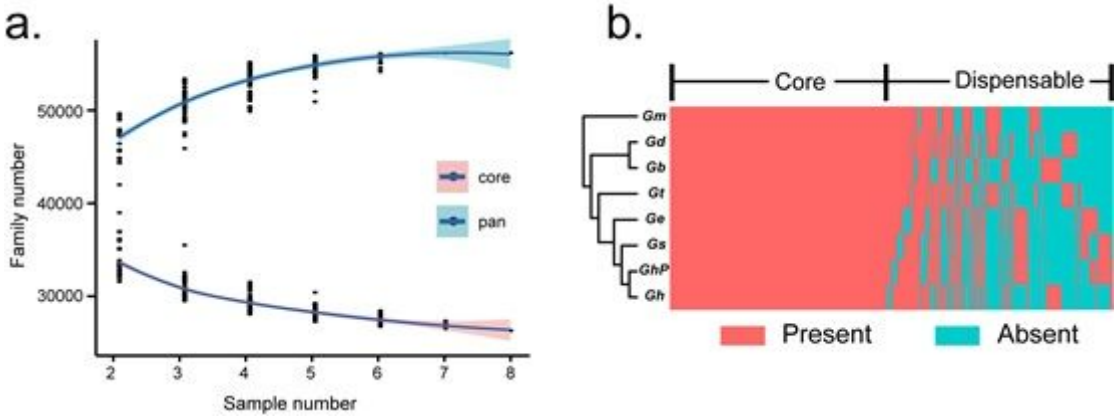
102. Tang S, Tang Z, Qiu L, Yang Z, Li G, Lang T, Zhu W, Zhang J, Fu S: **Developing New Oligo Probes to Distinguish Specific Chromosomal Segments and the A, B, D Genomes of Wheat (Triticum aestivum L.) Using ND-FISH.** *Front Plant Sci* 2018, **9:**1104.

# Table

**Table 1| Features of three tetraploid cotton assemblies.**

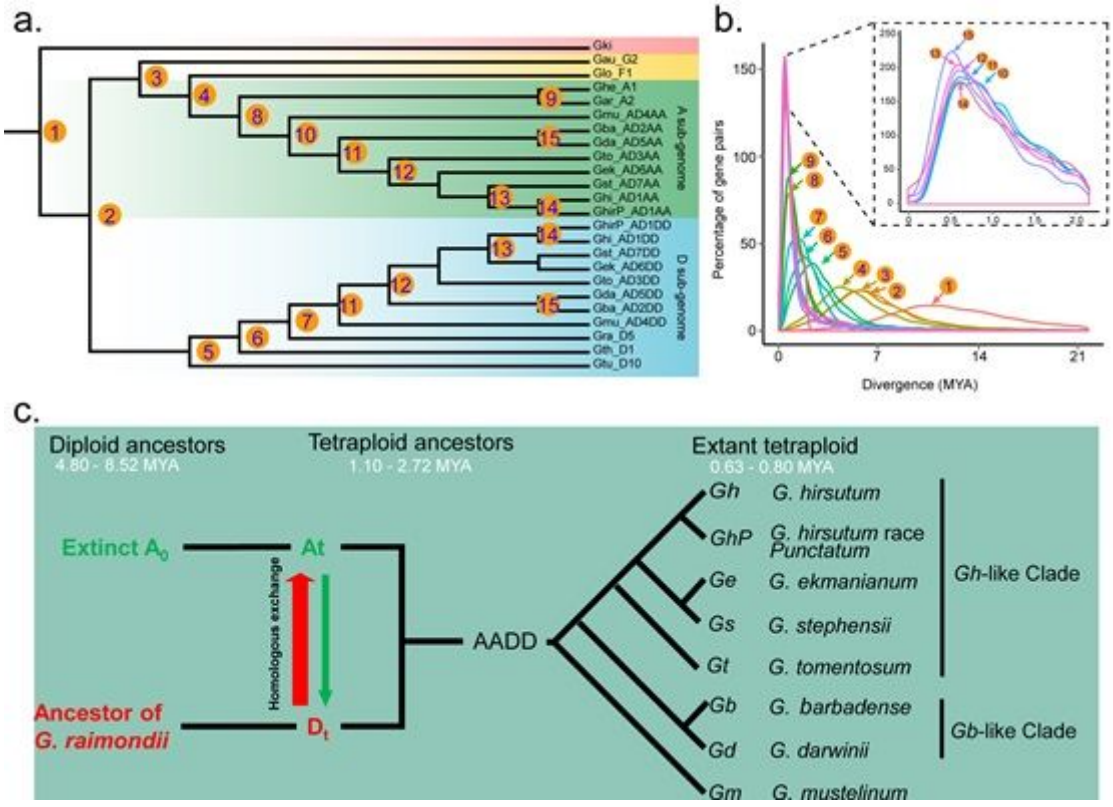
| Genomic features | *Ge* | *Gs* | *GhP* |
|---|---|---|---|
| Assembly | | | |
| Total length of scaffolds (Mb) | 2,341.87 | 2,291.84 | 2,292.48 |
| Total number of scaffolds | 160 | 243 | 277 |
| Scaffold N50 (Mb) | 108.06 | 108.2 | 106.96 |
| Total length of contigs (Mb) | 2,341.51 | 2,291.47 | 2,292.40 |
| Total number of contigs | 3,781 | 3,927 | 1,111 |
| Contig N50 (Mb) | 1.57 | 1.23 | 11.49 |
| Gap counts | 3,621 | 3,684 | 834 |
| Gap length (Mb) | 0.36 | 0.37 | 0.08 |
| Pseudo-chromosomes length（Mb） | 2,337.03 | 2,272.89 | 2,283.07 |
| Annotation | | | |
| Percentage of repeat sequences（%） | 64.86% | 63.01% | 64.89% |
| Number of genes | 74,178 | 74,970 | 74,520 |
| Genes in pseudochromosomes | 74,038 | 73,324 | 74,283 |
| Complete BUSCOs | 95.50% | 97.10% | 95.40% |

# Figures



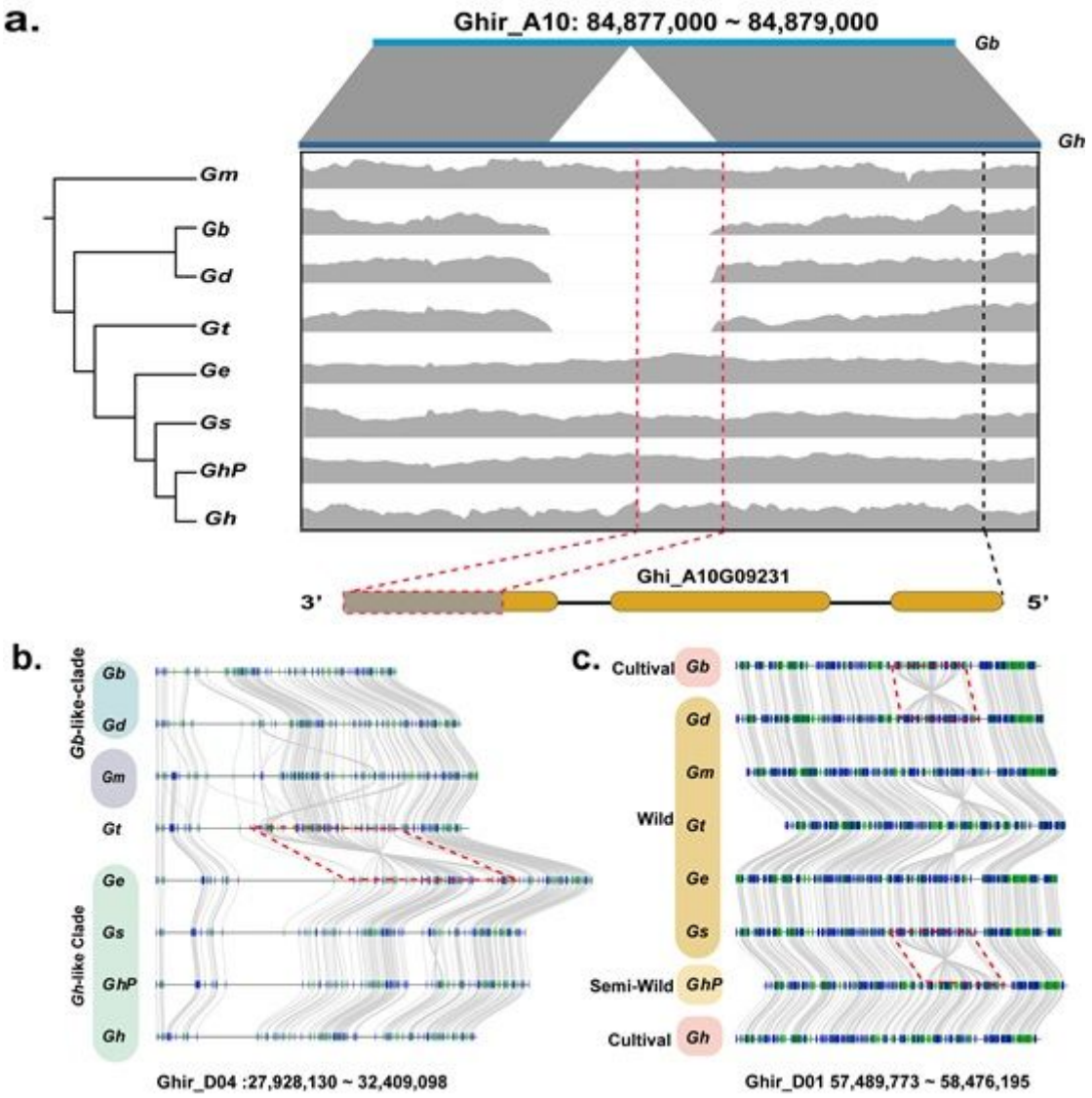

## Figure 1

Pangenome 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.



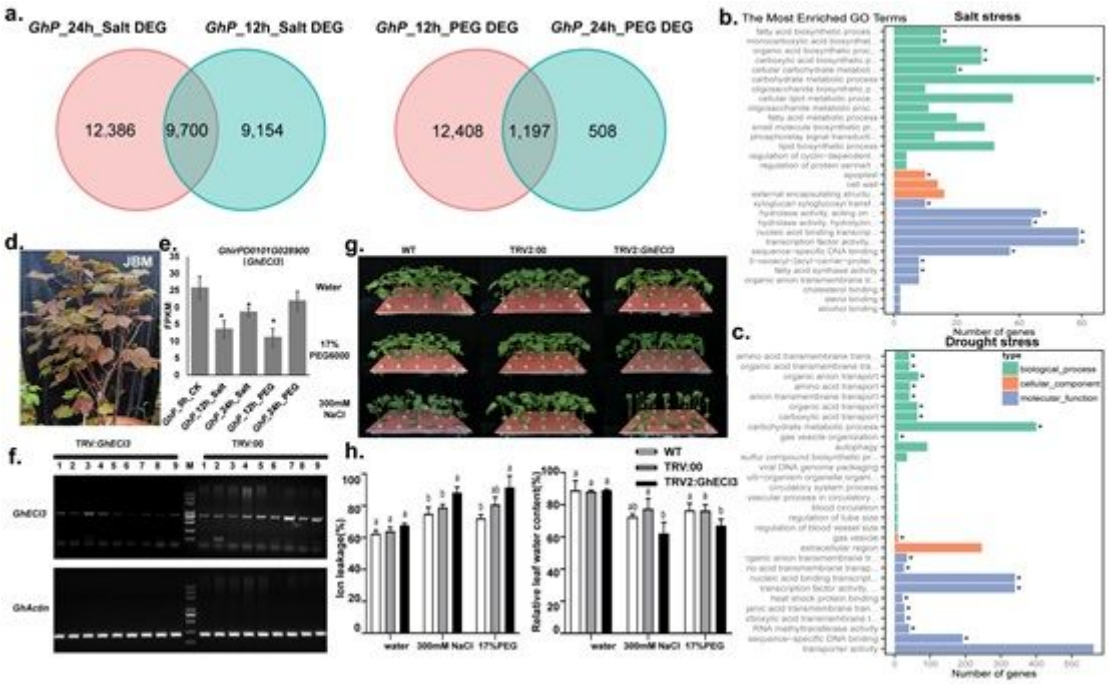

## Figure 2

Phylogenetic analysis of the Gossypium genomes. a. Maximum likelihood tree inferred using Gossypioides 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 A0 and ancestor of G. raimondii.




## Figure 3

Structure variations (SVs) in tetraploid cotton genomes. a. A deletion region between Gh and Gb. Top panel, comparison between Gh and Gb sequences show a ~500-bp fragment deletion. Coverage of Gh genome by eight tetraploid Gossypium genome reads and gene structure of Ghi_A10G09231 are shown at bottom; the deletion region is outlined in red. Evolutionary relationships are shown in the tree to the left. (b.) A 4.5 Mb inversion within the lineage leading to Ge-Gs-GhP-Gh. (c.) A 980 kb inversion shared by cultivated Gb and Gh relative to their wild progenitors.




## Figure 4

Abiotic stress adaption of GhP. a, DEGs in salt and drought. GhP_12h/24h_Salt/PEG represents the time point (12h or 24h) after salt (300mM NaCl) or drought (17% PEG) treatment. b and c, the most enriched GO terms between 12h and 24h, enrichment analysis under salt and drought stress, respectively. d, the mature plants of GhP. e, expression level of GhECI3. f, decreased GhECI3 expression in VIGS plants. g, decreased salt and drought tolerance of GhECI3 silenced plants. h and i show ion leakage and relative leaf water content, respectively. ANOVA analysis was performed with the standard t-test, with least significant difference (LSD) used for multiple comparisons.

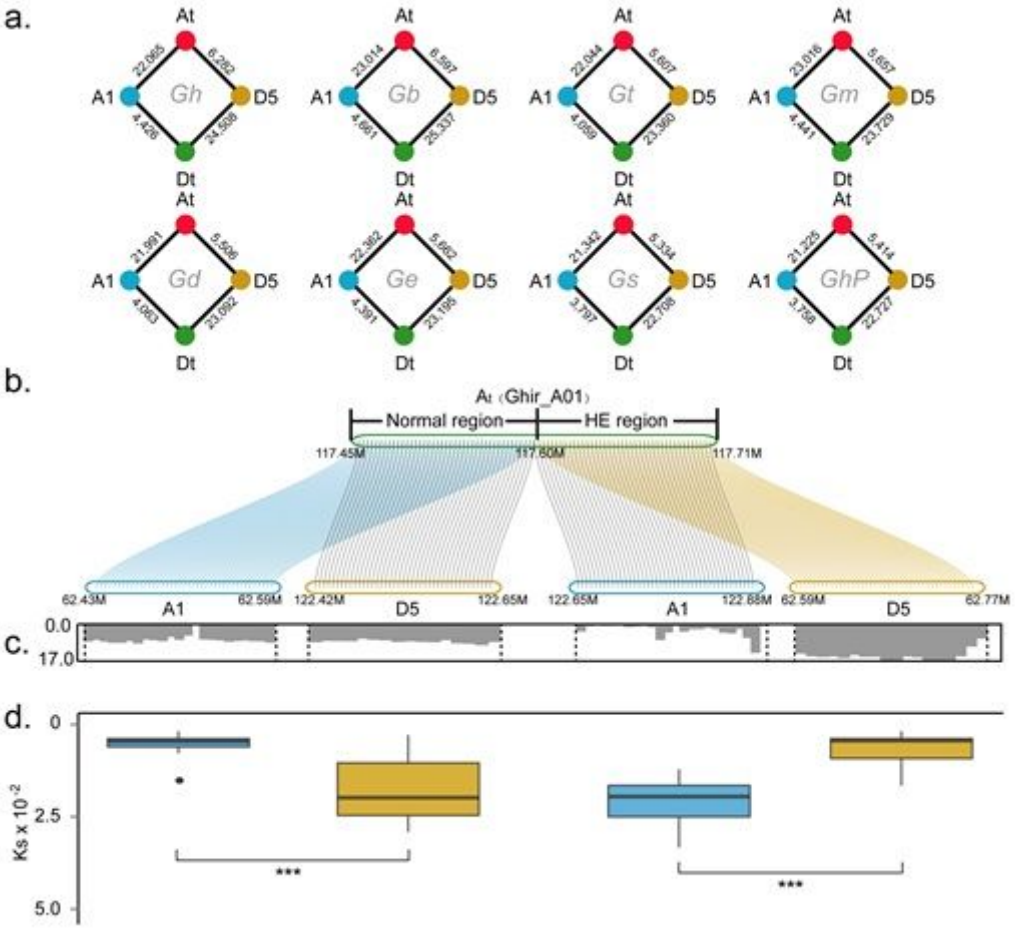



**Figure 5**

Homologous exchange (HE) in tetraploid cotton genomes. a. Statistics of gene conservation. The number at the top of each solid line indicates the ortholog number of genes between a tetraploid subgenome and its corresponding diploid ancestor. In the box of each drawing is the name of tetraploid genome. b. An example of HE from Dt to At in Gh genome. The blue line indicates that the segment is from the A1 genome, and the yellow line indicates that the corresponding segment is from the D5 genome. c. Coverage depths of reads of tetraploid genome aligned on the parental A1 and D5 genomes. d. Comparison of similarity between HE fragment and diploid homologous fragment. The lower panel shows the Ks score value distribution for syntenic blocks, which indicates HE in the tetraploid cotton.

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigures.docx
- SupplementaryTables.xlsx