


RESEARCH ARTICLE

Open Access



# Cotton D genome assemblies built with long-read data unveil mechanisms of centromere evolution and stress tolerance divergence

Zhaoen Yang<sup>1,2†</sup>, Xiaoyang Ge<sup>1,2†</sup>, Weinan Li<sup>3†</sup>, Yuying Jin<sup>2</sup>, Lisen Liu<sup>2</sup>, Wei Hu<sup>2</sup>, Fuyan Liu<sup>6</sup>, Yanli Chen<sup>2</sup>, Shaoliang Peng<sup>3,4,5\*</sup> and Fuguang Li<sup>1,2\*</sup> 

## Abstract

**Background:** Many of genome features which could help unravel the often complex post-speciation evolution of closely related species are obscured because of their location in chromosomal regions difficult to accurately characterize using standard genome analysis methods, including centromeres and repeat regions.

**Results:** Here, we analyze the genome evolution and diversification of two recently diverged sister cotton species based on nanopore long-read sequence assemblies and Hi-C 3D genome data. Although D genomes are conserved in gene content, they have diversified in gene order, gene structure, gene family diversification, 3D chromatin structure, long-range regulation, and stress-related traits. Inversions predominate among D genome rearrangements. Our results support roles for 5mC and 6mA in gene activation, and 3D chromatin analysis showed that diversification in proximal-vs-distal regulatory-region interactions shape the regulation of defense-related-gene expression. Using a newly developed method, we accurately positioned cotton centromeres and found that these regions have undergone obviously more rapid evolution relative to chromosome arms. We also discovered a cotton-specific LTR class that clarifies evolutionary trajectories among diverse cotton species and identified genetic networks underlying the *Verticillium* tolerance of *Gossypium thurberi* (e.g., SA signaling) and salt-stress tolerance of *Gossypium davidsonii* (e.g., ethylene biosynthesis). Finally, overexpression of *G. thurberi* genes in upland cotton demonstrated how wild cottons can be exploited for crop improvement.

**Conclusions:** Our study substantially deepens understanding about how centromeres have developed and evolutionarily impacted the divergence among closely related cotton species and reveals genes and 3D genome structures which can guide basic investigations and applied efforts to improve crops.

**Keywords:** *G. thurberi*, *G. davidsonii*, Salt tolerance, *Verticillium wilt*, Structural variation, Long-range interactions, Hi-C, 3D genome, *Gossypium*

\* Correspondence: [slpeng@hnu.edu.cn](mailto:slpeng@hnu.edu.cn); [aylifug@caas.cn](mailto:aylifug@caas.cn)

<sup>†</sup>Zhaoen Yang, Xiaoyang Ge and Weinan Li contributed equally to this work.

<sup>3</sup>College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China

<sup>1</sup>Zhengzhou Research Base, State Key Laboratory of Cotton Biology, Zhengzhou University, Zhengzhou 450001, China

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

Modern cultivated cotton has narrow genetic diversity, a situation which limits the improvement potential of these species [1, 2]. Besides the four cultivated species, there are more than 45 wild cotton species, and these have been grouped into nine genomic types (A–G plus K for diploids; AD for tetraploids) based on their kinship; these wild cotton species represent important resources for cotton breeding and the study of cotton evolution and domestication [3–6]. The diploid D genome type comprises 13 species, distributed from Southwest Mexico to Arizona, with additional disjunct species distributions in Peru and the Galapagos Islands [7].

Even though none of the D diploid species produce commercial fibers, the diploid D genome is known as the donor of the D subgenome in wild and domesticated allotetraploid cotton, and the D genome harbors potentially useful genes for improving fiber quality, disease and pest resistance, and cytoplasmic male sterility, as well as drought and salt tolerance in domesticated cotton [8, 9]. Because of their close relationships to the agronomically important cultivated cotton, the diversity, distribution, phylogenetic relationships, and taxonomy of the D genome wild cotton species have attracted scientific interest [7, 10].

Genomics research about D genomes was substantially advanced by the sequencing and de novo assembly of genomes for *G. raimondii* (D<sub>5</sub>) and *Gossypium turneri* (D<sub>10</sub>), yet genomic information for most D genome species remains unavailable [11, 12]. *G. thurberi* (D<sub>1</sub>) and *G. davidsonii* (D<sub>3</sub>) have the same number of chromosomes and similar content of genes with the closely related *G. raimondii* and *G. turneri* [7]. However, the phylogenetic and genetic data—as well as the plant classifications recognized by early taxonomists—support that *G. thurberi* and *G. davidsonii* are genetically distinct from *G. raimondii* and *G. turneri* [7]. Moreover, they also showed different phenotypic characters: compared with *G. raimondii*, *G. thurberi* is more tolerant to *Verticillium* wilt and *G. davidsonii* is more tolerant to salt.

Functional impacts from centromeres have been appreciated for more than 130 years, and although centromeres have been characterized using both cytological and genetic approaches, elucidating the molecular basis through which centromere exert their functional impacts is a central, ongoing pursuit in molecular biology research [13]. We know that centromeres are functionally conserved across eukaryotes, for example helping to ensure faithful transmission of the genome during cell division, but centromeres are often poorly represented in the genome assemblies. This poor representation reflects the highly repetitive sequences comprising centromeres, which has made them the most technologically challenging genome regions to assemble, particularly when

using short-read sequencing data. Indeed, for many species, centromere positions on chromosomes are still determined based on phylogenetic analyses or chromatin immunoprecipitation, methods which are both laborious and indirect [14]. The inability to accurately assemble centromeres in genome assemblies has limited our understanding of the functional mechanisms and evolutionary histories of these highly impactful genomic structures. And the capacity to resolve centromeres has been one of the major application cases for the introduction of long-read sequencing technologies into genomics research.

Here, we report high-quality genomes for *G. thurberi* and *G. davidsonii* that were assembled based on nanopore long reads and high-throughput chromosome conformation capture (Hi-C) technology, thereby substantially improving the quality and utility of the genomic resources available for research in this important cash crop. Using these high-quality genomes, we performed genome-wide comparative studies that revealed the contrasting features among *G. thurberi*, *G. davidsonii*, and *G. raimondii*, including for example chromosomal reconstruction analysis of species divergence, gene family expansion, gene-order and structural variations, methylation features, and long-range interactions between proximal and distal regulatory regions. Some findings from these analyses include the observation that genes with chromosomal interactions have higher expression levels than those without interactions, the finding that the relatively low levels of 5mC and 6mA in A compartments and at TAD boundaries may contribute to the activation of nearby genes, and a demonstration that recent *Gypsy* LTR expansion has driven the substantial divergence in orthologous centromere sequences among these closely related species. We also found that enhanced SA signaling and ethylene biosynthesis contribute to the respective abilities of *G. thurberi* and *G. davidsonii* to cope with biotic (*Verticillium dahliae*) and abiotic (salinity) stress.

## Results

### Genome sequencing, assembly, and annotation

We assembled the genomes of *G. thurberi* and *G. davidsonii* using data from both the nanopore long-read and the Hi-C short-read technologies. We produced 114.3 Gb and 108.3 Gb clean reads, respectively, for *G. thurberi* (~146×) and *G. davidsonii* (~135×) using the Nanopore platform (Additional file 1: Table S1–S2). After correction using the Illumina short reads, we generated a *G. thurberi* genome of 779.6 Mb with a contig N50 of 24.7 Mb; the corresponding values for *G. davidsonii* were 801.2 Mb and 26.8 Mb (Table 1 and Additional file 1: Table S3); the sequence continuities are

significantly improved for both species as compared with other recently reported genome assemblies [15, 16].

Using 284 million and 280 million valid Hi-C interaction pairs for the *G. thurberi* and *G. davidsonii* genomes, respectively (Additional file 1: Table S4), we anchored and oriented 777.2 and 799.2 Mb of the assembly onto 13 pseudochromosomes of *G. thurberi* and *G. davidsonii* respectively (Additional file 2: Fig. S1-S2), which represented more than 99.7% of the total assembly, indicating that our new assemblies reached a reference grade for quality. As an indication of the improved contiguity, the contig length for our *G. thurberi* genome represents a 940-fold increase compared to previously published *G. thurberi* sequences (24.7 Mb versus 0.026 Mb) [7], and our *G. thurberi* genome has a 3750-fold reduction in fragmentation (74 versus 277,903). Similarly, there was an 836-fold increase for *G. davidsonii* genome contig length (26.8 Mb versus 0.032 Mb) and 5150-fold reduction in fragmentation (104 versus 535,698). Moreover, the total assembly length and gene annotation number were all higher for our *G. thurberi* and *G. davidsonii* genome assemblies as compared to the recently reported *G. thurberi* and *G. davidsonii* genome resources [7]. Approximately 58.0% and 58.6% of the assembly sequences were annotated as repetitive sequences in the *G. thurberi* and *G. davidsonii* assemblies, respectively (Additional file 1: Table S5).

We next evaluated the assembly completeness by aligning the 192 and 212 million paired-end Illumina short reads against the *G. thurberi* and *G. davidsonii* genome assemblies and BUSCO [17] analysis, both methods showed that both assemblies are of high quality (Additional file 1: Table S6-S7 and Additional file 2: Fig. S3).

**Genomic diversity among six D genomes**

Our generation of high-quality genome assemblies for *G. thurberi* and *G. davidsonii* provides an opportunity to compare different D genome species that shared a common ancestor, potentially helping identify the post-divergence genomic rearrangements in cotton. The overall collinearities between the two newly assembly genomes are largely conserved, as supported by more than

78% of *G. thurberi* genome matching in one-to-one syntenic blocks with 80.6 % of the *G. davidsonii* genome. Similarly, we found approximately 78% of the *G. thurberi* genome matched in one-to-one syntenic blocks with ~ 81% of the *G. raimondii* genome (Additional file 1: Table S8). And ~ 77% *G. davidsonii* genome matched in one-to-one syntenic blocks with ~ 83% of the *G. raimondii* genome. Our previous study showed that ~ 86% of the *G. raimondii* genome matched in one-to-one syntenic blocks with the D subgenome of *Gossypium hirsutum* (Gh\_D<sub>t1</sub>), confirming that *G. raimondii* is a plausible donor species of allotetraploid cotton species [4].

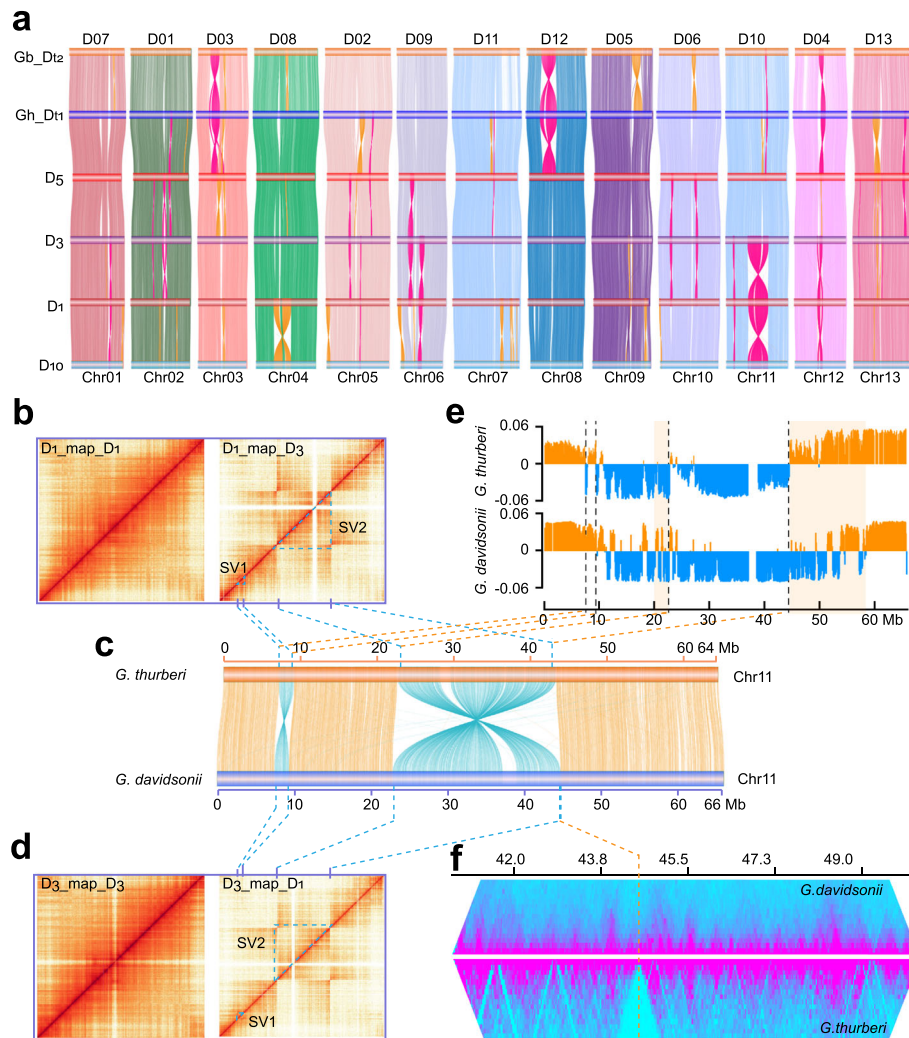
We found inversions are the major rearrangement type among the different D genomes. The inversions between the two new assemblies span approximately 59.6 Mb in *G. thurberi*, a level similar to a previously reported comparison between *G. raimondii* and the TM-1 D subgenome [4]. Of particular note, we detected a large inversion on Chr11 between the *G. thurberi* and *G. davidsonii* occupying 20.4 Mb; note, this was confirmed by mapping the Hi-C data for one accession against to the genome of the other, and *vice versa* (Fig. 1a-d and Additional file 2: Fig. S4-S5). Enlargements from the heatmaps revealed discontinuous signals for these inversions (in the region marked by the color triangle in Fig. 1b).

Our finding that *G. davidsonii*, *G. turneri*, and *G. raimondii* share a conserved syntenic relationship for the large Chr11 inverted region supports that this Chr11 inversion is specific to *G. thurberi*. Further, we detected that *G. thurberi* Chr11 exhibits extensive B-to-A compartment switching specifically in a region neighbor the right breakpoint of the large inversion (Fig. 1e). And we also found that, relative to *G. davidsonii*, the topologically associating domains (TAD) were obviously extensively reorganized near the breakpoints of *G. thurberi* Chr11 breakpoints (Fig. 1f). We analyzed the conserved and switched A-B compartments in the inverted regions and at the whole-genome level. We found 39 conserved and 26 switched A-B compartments in the inverted regions, and the corresponding values for the whole genomes were 1045 and 532. Chi-square tests suggested that there was no bias towards A-B compartment

**Table 1** Global statistical comparison of *G. thurberi* and *G. davidsonii* genome

Category	<i>G. thurberi</i>					<i>G. davidsonii</i>				
	Numbers	N50 (Mb)	Longest (Mb)	Size (Mb)	Percentage of assembly	Numbers	N50 (Mb)	Longest (Mb)	Size (Mb)	Percentage of assembly
Contigs	74	24.7	49.9	779.6	100	104	26.8	47.4	801.2	100
Anchored and oriented	63	24.7	49.9	777.3	99.7	90	26.8	47.4	799.1	99.7
Gene annotated	41,316	NA	NA	111.9	14.4	41,471	NA	NA	113.9	12.5
Repeat sequence	NA	NA	NA	451.8	58.0	NA	NA	NA	469.4	58.6

NA not applicable



**Fig. 1.** Characterization of genomic variation among different D genomes. **a** Genome comparison of among *G. barbadense* (D subgenome, Gb\_D<sub>12</sub>), *G. hirsutum* (D subgenome, Gh\_D<sub>11</sub>), *G. raimondii* (D<sub>3</sub>), *G. davidsonii* (D<sub>3</sub>), *G. thurberi* (D<sub>1</sub>), and *G. turneri* (D<sub>10</sub>). The inversions are marked in orange and magenta. **b** Identification of a large inversion on Chr11 between *G. thurberi* and *G. davidsonii*. The panel shows chromatin interaction heat maps including *G. thurberi* Hi-C data mapping *G. thurberi* (D<sub>1</sub>\_map\_D<sub>1</sub>) and *G. davidsonii* Hi-C data mapping *G. thurberi* (D<sub>1</sub>\_map\_D<sub>3</sub>). The triangle marks the inversions in the heat maps. **c** Genomic comparison between *G. thurberi* and *G. davidsonii* on Chr11. **d** The panel shows chromatin interaction heat maps including *G. davidsonii* Hi-C data mapping *G. davidsonii* (D<sub>3</sub>\_map\_D<sub>3</sub>) and *G. thurberi* Hi-C data mapping *G. davidsonii* (D<sub>1</sub>\_map\_D<sub>3</sub>). The triangle marks the inversions in the heat maps. **e** A/B compartments in Chr11; orange represents the A compartments and blue represents the B compartments. The transparent boxes indicate A-B compartment switching regions. **f** TAD heatmap around the right breakpoint of the large inversion on Chr11

switching in inverted regions (chi-square test,  $P = 0.2664$ ). In contrast, we detected and obviously elevated proportion of reorganized TAD boundaries near the breakpoints (70 out of 190) when compared to the whole genome (1143 out of 6184) (chi-square test,  $P < 0.0001$ ) (Additional file 2: Fig. S6). These results offer empirical demonstrations showing that inversions in plant genomes can—in addition to their better understood impacts on one-dimensional linear genome sequences divergence—also drive divergence in TAD boundary formation.

In addition to Chr11, we also found some inversions from Chr01, Chr05, Chr06, and Chr12 that are specific

to *G. thurberi* because they are shared by *G. davidsonii* and *G. turneri* (Fig. 1a). Similarly, some inversions from *G. davidsonii* include inversions from Chr02, Chr05, Chr06, Chr10, and Chr13, which are shared by *G. raimondii* and *G. thurberi*. Furthermore, we observed that in *G. turneri* (D<sub>10</sub>), *G. hirsutum*, and *Gossypium barbadense* (Gb\_D<sub>12</sub>), most of inversions are species specific (Fig. 1a), indicating that such inversions have formed during species divergence; such structural rearrangements could have directly contributed genetic novelty that contributed to such divergence.

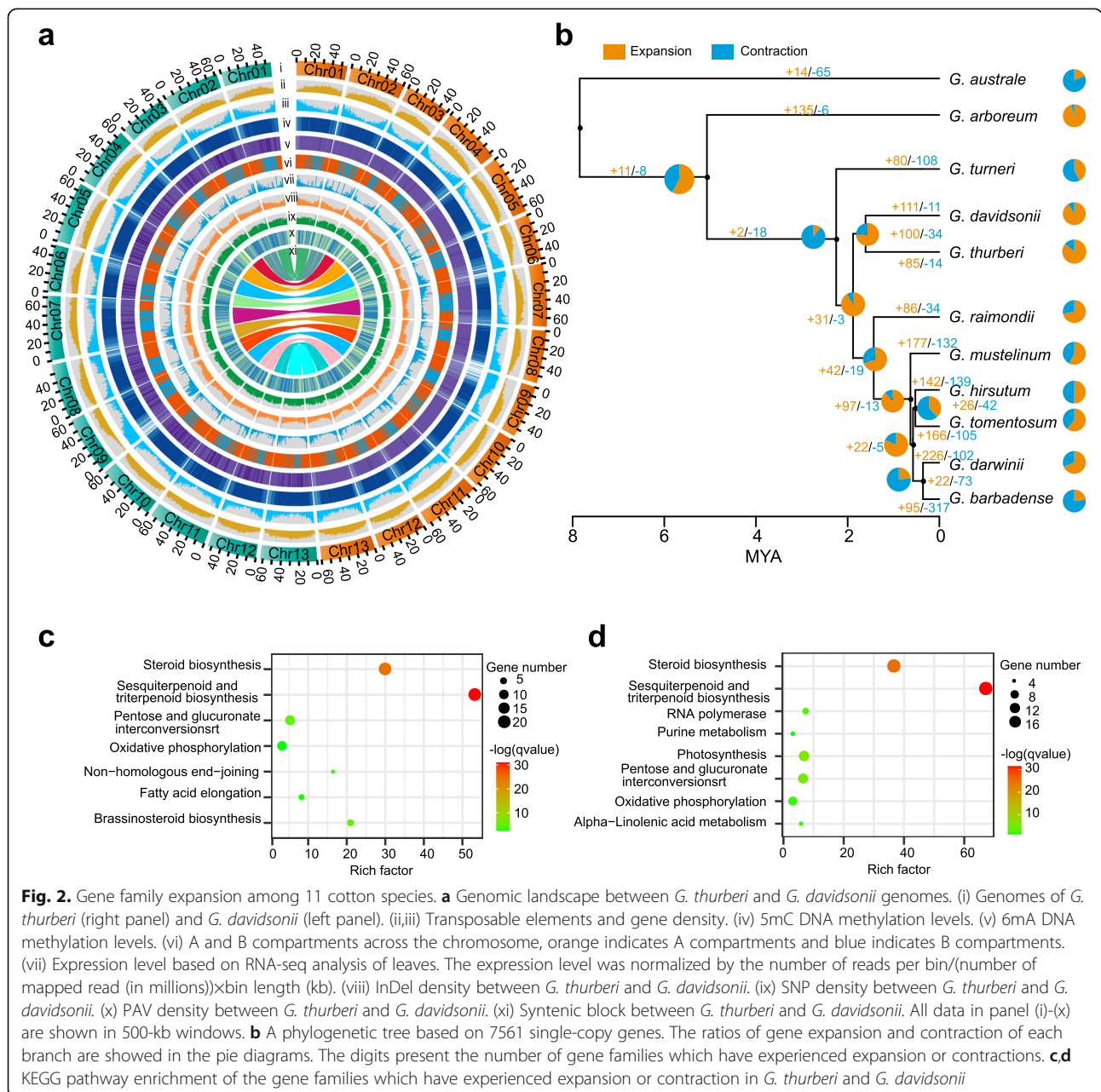
**Genomic landscapes of *G. thurberi* and *G. davidsonii***

As with most genomes, the *G. thurberi* and *G. davidsonii* sequences positioned near the telomere are enriched of coding genes while having a lower-than-average level of repeat sequences (Fig. 2a). Again as expected, the pericentromeric regions are enriched for repeat sequences but show a deficit for coding genes compared to the genome-wide average (Fig. 2a).

Our RNA-seq (coding and non-coding) expression profiling of *G. thurberi* and *G. davidsonii* (young leaves) showed that sequences in pericentromeric regions are expressed at generally lower levels compared to sequences in chromosome arms (Fig. 2a). We next

examined small variations (InDels and SNPs) between *G. thurberi* and *G. davidsonii*. The InDel densities showed a decreasing pattern, and the SNP densities showed an increasing tendency moving from the telomere region to the centromere region (Fig. 2a and Additional file 2: Fig. S7).

We next detected presence/absence variations (PAVs) and identified a total of 14,401 of *G. thurberi*-specific genomic PAVs and 15,684 of *G. davidsonii*-specific genomic PAVs, occupying 39.5 Mb and 52.0 Mb in *G. thurberi* and *G. davidsonii* genomes. The PAVs are evenly distributed across the chromosomes, with most of the



**Fig. 2.** Gene family expansion among 11 cotton species. **a** Genomic landscape between *G. thurberi* and *G. davidsonii* genomes. (i) Genomes of *G. thurberi* (right panel) and *G. davidsonii* (left panel). (ii,iii) Transposable elements and gene density. (iv) 5mC DNA methylation levels. (v) 6mA DNA methylation levels. (vi) A and B compartments across the chromosome, orange indicates A compartments and blue indicates B compartments. (vii) Expression level based on RNA-seq analysis of leaves. The expression level was normalized by the number of reads per bin/(number of mapped read (in millions))xbin length (kb). (viii) InDel density between *G. thurberi* and *G. davidsonii*. (ix) SNP density between *G. thurberi* and *G. davidsonii*. (x) PAV density between *G. thurberi* and *G. davidsonii*. (xi) Syntenic block between *G. thurberi* and *G. davidsonii*. All data in panel (i)-(xi) are shown in 500-kb windows. **b** A phylogenetic tree based on 7561 single-copy genes. The ratios of gene expansion and contraction of each branch are showed in the pie diagrams. The digits present the number of gene families which have experienced expansion or contractions. **c,d** KEGG pathway enrichment of the gene families which have experienced expansion or contraction in *G. thurberi* and *G. davidsonii*

PAVs being shorter than 10 kb (Fig. 2a and Additional file 2: Fig. S8a).

A total of 490 and 570 PAV-localized genes were identified as *G. thurberi*- or *G. davidsonii*-specific genes. Approximately 39.6% and 37.4% of the PAV genes from *G. thurberi* and *G. davidsonii* had apparent orthologs from at least one of the three other *Gossypium* species, underscoring that a relatively small proportion of these PAV genes were present in the ancestral genome (Additional file 2: Fig. S8b). PAV genes without obvious orthologs in the examined *Gossypium* species are likely to have arisen during the divergence and may represent sources of impactful genes that have contributed to the speciation and presently adapted characteristics of *G. thurberi* and *G. davidsonii*.

#### Evolution within and between eleven cotton genomes

We also compared new coding genes from the new assemblies with *Gossypium arboreum* ( $A_2$ ), *Gossypium australe* ( $G_2$ ), *G. raimondii* ( $D_5$ ), *G. turneri* ( $D_{10}$ ), and the D subgenomes of the five allotetraploid cotton species (*G. hirsutum*, *G. barbadense*, *Gossypium tomentosum*, *Gossypium mustelinum*, and *Gossypium darwinii*). Our phylogenetic tree supports a monophyletic origin for the allotetraploid species that was likely derived from a hybridization between *G. raimondii* and an A genome species (Fig. 2b). A total of 35,454 orthologous groups were identified through orthoMCL, and as expected, the *G. australe* ( $G_2$ ) and *G. arboreum* ( $A_2$ ) has more unique genes than those of D genome species, because the genomic divergences are more significant in diverse chromosomal groups than within a single group (Additional file 2: Fig. S9a). GO analysis revealed enrichment for “DNA recombination,” “DNA integration,” and “DNA metabolic process” among the unique gene sets for *G. thurberi* and *G. davidsonii* (Additional file 2: Fig. S9b-c).

Using *G. arboreum* and four D genome species (*G. thurberi*, *G. davidsonii*, *G. raimondii*, and *G. turneri*), we evaluated the divergence times between the diploid A genome and four D genome species and found they apparently diverged between 5.07 and 5.13 MYA, and the four D genomes diverged between 1.51 and 2.04 MYA (Additional file 2: Fig. S10). Within the D genome clade, the greatest extents of divergence were detected between *G. turneri* and the other 3 species, then the followed divergence was between *G. raimondii* and the other 3 species, and the most recent divergence was between *G. thurberi* and *G. davidsonii* (Fig. 2a).

We next used CAFE (Computational Analysis of gene Family Evolution) to estimate gene family expansions and contractions among the 23,825 ortholog groups, which revealed that 8 out of the 11 tested species have experienced more gene family expansions than gene

family contractions ( $P < 0.05$ ) (Fig. 2b). This is informative when considered against gene family dynamics known for the D subgenomes of allotetraploid cottons: our finding that a relatively higher proportion of species-specific gene families have experienced expansion or contractions in diploid D genome species compared with gene family dynamics in D subgenomes support that this form of genome divergence is less active in the D subgenomes than in the D genome species (Additional file 2: Fig. S11).

We detected that *G. thurberi* has experienced expansion for genes related to steroid biosynthesis and brassinosteroid biosynthesis, as well as for genes encoding pectinesterase enzymes (Additional file 2: Fig. S12). Given the reported roles of these biochemical pathways and enzymes in diverse stress tolerance responses, perhaps such expansion has contributed to the previously reported *Verticillium dahliae* resistance of *G. thurberi* [18]. Enriched genes specific to *G. davidsonii* included genes which function in photosynthesis and oxidative phosphorylation pathways, in the photosystem I reaction center (PsaB), and in the photosystem II reaction center (psbD and psbE) are enriched in *G. davidsonii* (Additional file 2: Fig. S13), results clearly suggesting that the potential for differential photosynthetic capacities in *G. davidsonii*.

#### Epigenetic modifications variations in 3D structure

Both PacBio and Nanopore can distinguish modified bases from standard nucleotide bases in plants [19, 20]. However, the accuracy of SMRT sequencing for detecting DNA methylation is known to be heavily affected by the sequence coverage [21, 22]. We used the nanopore data to analyze the global landscape of epigenetic modifications on chromosomes. The global N6-methyldeoxyadenine (6mA) level is approximate 1.1% of all adenines for *G. thurberi* and 1.3% for *G. davidsonii*, these proportions are much higher than previous reports about *G. hirsutum* and *G. barbadense* that were based on PacBio sequencing data [19]. For both *G. thurberi* and *G. davidsonii*, the 6mA distribution is uneven across the chromosomes, for example exhibiting enrichment at both the middle regions of chromosome arms and in pericentromeric regions (Fig. 2a), findings supporting the proposal from a rice study that the genomic distribution of 6mA is not random [23]. A comparison of the *G. davidsonii* genome methylation frequencies generated using Nanopore technology with the methylation frequencies obtained through whole-genome bisulfite sequencing technology showed an excellent correlation between the two methods ( $R = 0.88$ ). Among the three types of methylation (CHG, CG, and CHH), CHG showed the highest correlation (0.95), followed by CG (0.92) and CHH ( $R = 0.77$ ) (Additional file 2: Fig. S14).

The chromosome can be experimentally delineated into open (A) or closed (B) compartments, and these A/B compartments can be further divided into smaller TADs. We found that A compartments tended to cluster at chromosome arms, while B compartments tended to cluster near pericentromeric regions (Fig. 2a). Approximately 41.5% of the *G. thurberi* genome belongs to A compartments; this was 42.3% for *G. davidsonii* and 42.0% for *G. raimondii* genome. Note that these A/B compartment ratios are similar with ratios previously reported for allotetraploid D subgenomes [24].

We further evaluated the epigenetic features in the A/B compartments for *G. thurberi* and *G. davidsonii* in an analysis using 100-kb windows. For both the *G. thurberi* and *G. davidsonii* genomes, the gene densities were much higher in the A compartments than the B compartments (Fig. 3c). Further, it was intriguing to observe that the levels of both 5mC (CG, CHH, and CHG) and 6mA were significantly lower in A compartments than in B compartments (Fig. 3a). Similarly, the TE content was much lower in A compartments as compared to B compartments (Fig. 3b).

We also analyzed epigenetic modifications around TAD boundaries and found that chromatin surrounding the TAD boundaries in both of the examined cotton species had relatively lower levels of 5mC (CG, CHG, and CHH) and 6mA compared against randomly sampled genomic regions (Fig. 3d). Notably, there is enrichment for ORF sequences at TAD boundaries, suggesting that epigenetic modifications may apparently contribute to the differential activation of genes positioned at TAD boundaries.

To check for higher-order structural variations possibly related to the divergence of D genome species, we compared 3D structures among *G. thurberi*, *G. davidsonii*, and *G. raimondii*. Specifically, we constructed chromatin interaction maps for *G. thurberi*, *G. davidsonii*, and *G. raimondii* at 50 kb resolution, and as expected, the frequency of intra-chromosomal interactions displayed a rapid decrease with extended linear distance (Additional file 2: Fig. S15-S17). This analysis revealed strong rewiring of chromatin interactions in the inverted regions, consistent with a model of distinct territories formed by individual chromosome arms (Additional file 2: Fig. S18). For instance, *G. thurberi* carrying a pericentromeric inversion on Chr11 showed preferential interactions between these loci when present on the same chromosome arm (Additional file 2: Fig. S18).

We compare the organization of A/B compartments between the *G. thurberi* and *G. davidsonii* or between *G. davidsonii* and *G. raimondii*. A total of 57.8 Mb and 44.7 Mb in the *G. thurberi* and *G. raimondii* genomes represented apparent A-to-B compartment switching as compared with the compartment status data for *G.*

*davidsonii* (Fig. 3f). Similarly, a total of 28.9 Mb and 28.1 Mb of genome regions apparently represent B-to-A compartment switching between the *G. thurberi* and *G. raimondii* genomes (Fig. 3f), findings highlighting that B-to-A switching and A-to-B switching are uneven among the diploid D genomes. We also checked the potential for differential expression of genes located in the A/B switching regions: among 3189 A/B switching genes between *G. thurberi* and *G. davidsonii*, 556 were DEGs. Among 3670 A/B switching genes between *G. raimondii* and *G. davidsonii*, 613 were DEGs. These findings support the previous idea that only a small subset of genes are transcriptionally affected by compartment changes [25].

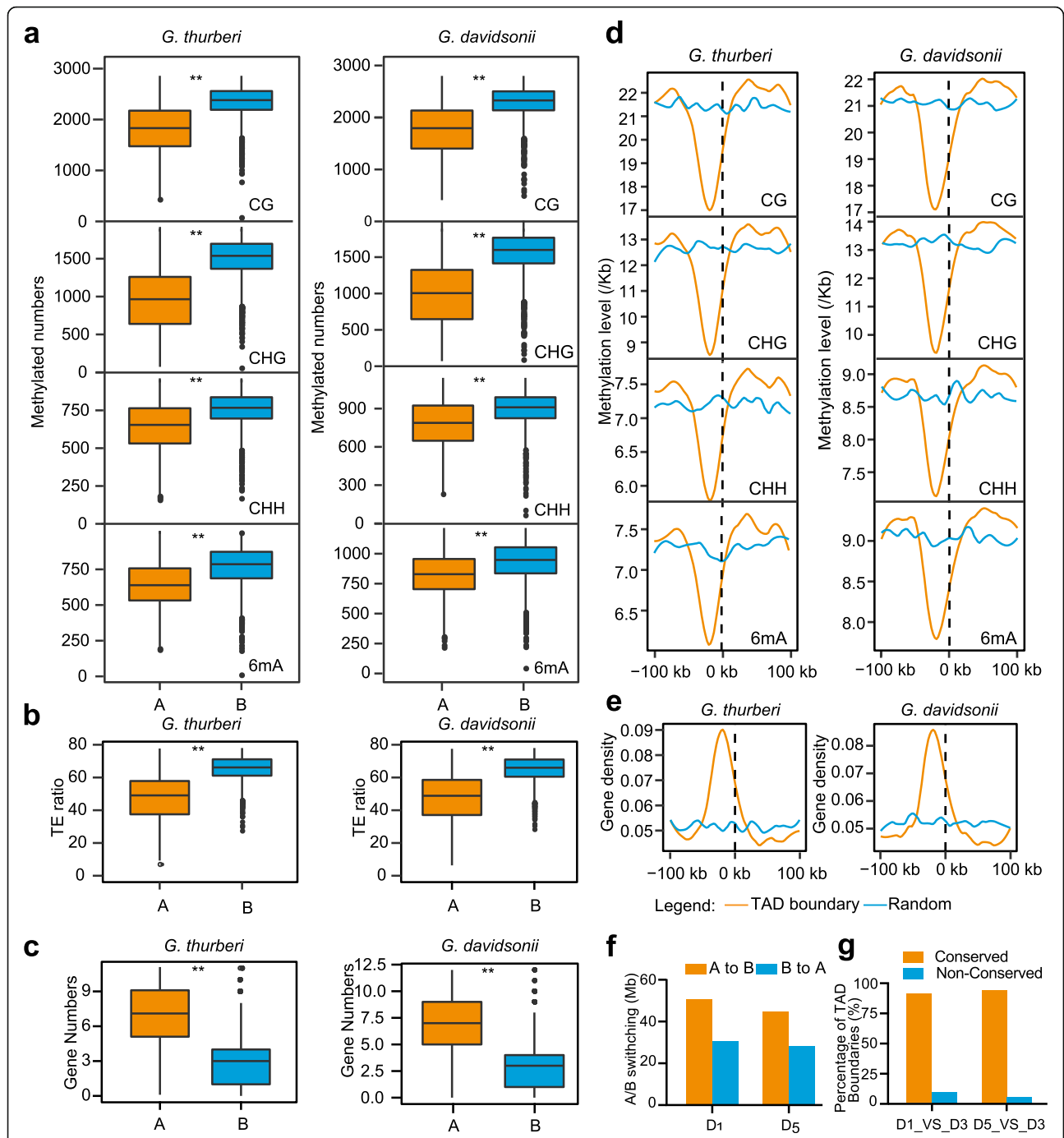
We next compared the TAD boundaries and found that more than 90% of the *G. thurberi* and *G. raimondii* TAD boundaries were conserved in *G. davidsonii* (Fig. 3g), indicating that the TAD boundaries have been relatively strongly conserved among sister species after divergence.

#### Long-range interactions in *G. thurberi* and in *G. davidsonii*

Long-range chromatin interactions functionally contribute to gene transcriptional regulation, but very little is known about 3D chromatin interactions in cotton. Seeking to characterize the pattern of long-range chromatin interactions, we conducted a genome-scale analysis and annotated the Hi-C peaks positioned within 2-kb upstream or 1-kb downstream of gene TSSs as “proximal Hi-C peaks” (P); all of the others were annotated as “distal Hi-C peaks” (D). We identified 22,328 P and 8304 D involved in long-range chromatin interactions in *G. thurberi*; *G. davidsonii* had 22,816 P and 8808 D involved in long-range chromatin interactions (Fig. 4a).

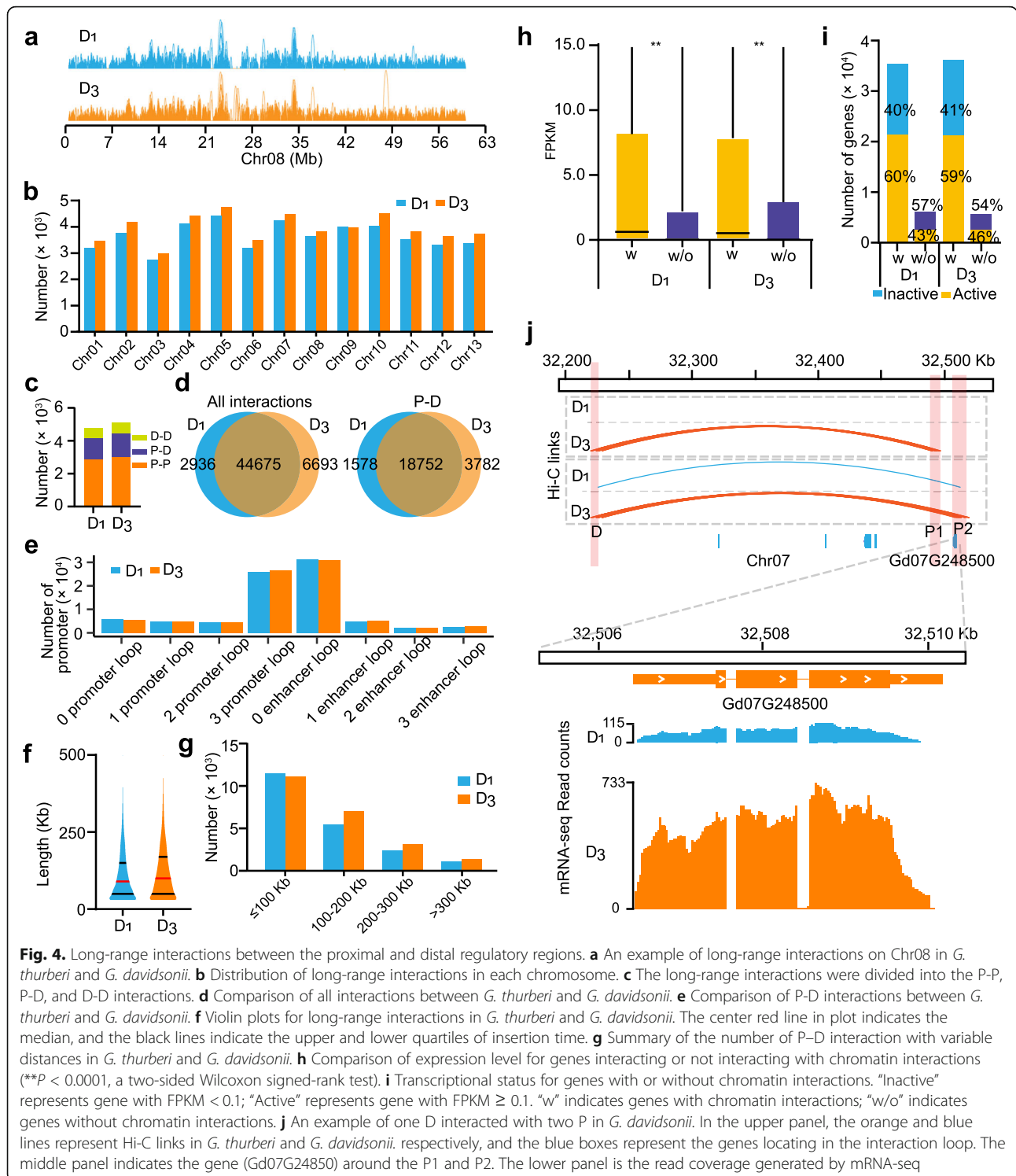
We also classified all of these interactions into three groups: proximal–proximal (P–P), proximal–distal (P–D), and distal–distal (D–D). A total of 47,604 and 51,367 intra-chromosomal interactions were identified for *G. thurberi* and *G. davidsonii*, respectively. Approximately 60% of these interactions were P–P interactions, followed by P–D (~30%) and D–D (~10%) (Fig. 4c). Comparison of the average number of formed loops is informative: we found that one D can form an average of 1.56 or 1.62 loops with P for *G. thurberi* and *G. davidsonii*, respectively; in contrast, one P can form an average of 1.34 or 1.31 loops with P. So, on average, D undergo more interactions than P, and it appears that genes regulated by D prefer to cluster together in the genome.

The number of interactions in each chromosome ranged from 2759 to 4417 in *G. thurberi* and 2999 to 4763 in *G. davidsonii* (Fig. 4b). There were 44,675 intra-chromosomal interactions were identified both in *G. thurberi* and *G. davidsonii*, while 2936 and 6693 interactions were specific to *G. thurberi* and *G. davidsonii*. We



**Fig. 3.** Methylation features of 3D chromatin. **a–c** Methylation level, TE ratio, and gene density in A and B compartments in *G. thurberi* and *G. davidsonii*. A two-sided Wilcoxon signed-rank indicates there were significant differences at  $***P < 0.001$ . **d** Methylation feature around TAD boundaries. The methylation levels in TAD boundaries (orange lines) flanking 100 kb was compared with those methylation levels in random genome regions (blue lines). The lines on the right side (0 to 100 kb) indicate TAD regions, and the lines on the left side (– 100 to 0 kb) indicate TAD regions when TADs were organized consecutively or non-TAD regions when one TAD was not closely adjacent to the others. **e** Gene density distribution around the TAD boundaries. The method for extracting genomic regions around boundaries was the same as that in panel **d**. **f** A-B compartment switching between *G. thurberi* (D<sub>1</sub>) and *G. davidsonii* (D<sub>3</sub>) or between *G. raimondii* (D<sub>5</sub>) and *G. davidsonii* (D<sub>3</sub>). **g** Comparison of TAD boundaries between *G. thurberi* and *G. davidsonii* (D<sub>1</sub>\_Vs\_D<sub>3</sub>) or *G. raimondii* and *G. davidsonii* (D<sub>5</sub>\_Vs\_D<sub>3</sub>)





found 27,531 P-P interactions, 18,752 P-D interactions, and 5465 D-D interactions were conserved between *G. thurberi* and *G. davidsonii*, whereas 1043 and 2597 P-P interactions, 1578 and 3782 P-D, and 761 and 1067 D-D interactions were respectively specific for *G. thurberi* and *G. davidsonii* (Fig. 4d). This result emphasizes that

only a small subset of intra-chromosomal interactions is divergent between cotton sister species.

Strikingly, more than 73% of the promoters of these cotton genomes have 3 or more P-P interactions, with most promoters having about 1 P-D (Fig. 4e). We found that the median length of intra-chromosomal

interactions were 90 kb and 100 kb for *G. thurberi* and *G. davidsonii*, respectively (Fig. 4f). Previous studies in human have showed that enhancers prefer to regulate nearby genes [26]. Most of the P-D interactions were within 100 kb for both species, and fewer than 6% of these interactions were larger than 300 kb (Fig. 4g).

We next generated the transcriptome datasets using mRNA-Seq of *G. thurberi* and *G. davidsonii* leaves to help experimentally unravel the relationships between chromatin interactions and the transcriptional activation of cotton genes. We compared the expression profiles of genes with or without chromosomal interactions and found that genes with chromatin interactions had relatively higher expression levels than those without interactions ( $P < 0.0001$ , Wilcoxon rank-sum test) (Fig. 4h). Although the chromatin interactions were captured by Hi-C, we found that ~40% of the genes with interactions not expressed or expressed very lowly (FPKM < 0.1) (Fig. 4i). However, between 43 and 46% of genes which had no chromatin interactions in either *G. thurberi* and *G. davidsonii* were expressed in leaves, a level slightly higher than from a report for such genes in an analysis of shoots and immature ears in maize [26] (Fig. 4i).

We then examined the intersection of the differentially expressed genes between *G. thurberi* and *G. davidsonii* and differential P-D interaction genes to explore the possible roles of enhancers on the gene expression. Among the genes with differential P-D interactions between *G. thurberi* and *G. davidsonii*, there 509 genes which significantly altered expression levels, and these genes exhibited enrichment for GO terms including “response to biotic stimulus” and “defense response” (Additional file 2: Fig. S19). An example of these genes is the homeobox gene *Gd07G248500*, which encodes an ortholog of *AtHB16*, which is known to regulate leaf development and photoperiod sensitivity in *Arabidopsis thaliana* [27]. We found that the promoter of *Gd07G248500* interacted with 2 D peaks in both *G. thurberi* and *G. davidsonii*, but the interaction intensities were stronger in *G. davidsonii* than those in *G. thurberi* (15.13\_vs\_0.06 and 12.94\_vs\_1.2), which may promote its expression in *G. davidsonii* leaves (Fig. 4j), a situation like the maize PSB1 that had a shoot-specific P-D interaction with a higher expression in shoot than that in immature ear [26].

**Table 2** Variations within genes between *G. thurberi* and *G. davidsonii* genomes

Variation type	Syntenic region		SV region	
	Number	Ratio	Number	Ratio
Structurally conserved genes <sup>a</sup>	24,094	75.58	729	66.03
Without amino acid substitutions <sup>b</sup>	2832	8.88	88	7.97
No DNA variation in CDS region	699	2.19	22	1.99
No DNA variation in CDS and intron region	292	0.92	10	0.91
No DNA variation in genic region <sup>c</sup>	4	0.01	0	0.00
Same sense mutation	2133	6.69	66	5.98
With amino acid changes <sup>d</sup>	21,262	66.70	641	58.06
With missense mutation in CDS	17,488	54.86	532	48.19
With 3n InDel in CDS	3774	11.84	109	9.87
Genes with large-effect mutations <sup>e</sup>	2915	9.14	126	11.41
With 3n ± 1 InDel in CDS	1035	3.25	32	2.90
Start-codon mutation	694	2.18	29	2.63
Stop-codon mutation	641	2.01	31	2.81
Splice-acceptor mutation	32	0.10	0	0.00
Splice-donor mutation	513	1.61	34	3.08
Genes with large structural variations <sup>f</sup>	4868	15.27	249	22.55
At least one CDS missing	4145	13.00	209	18.93
Total	32,981 <sup>g</sup>	100.00	1104 <sup>g</sup>	100.00

<sup>a</sup> Structurally conserved genes, including genes without amino acid substitutions (b) and with amino acid changes (d). <sup>b</sup> Genes without amino acid substitutions (no DNA variation in the CDS region or intron regions). <sup>c</sup> Genic regions including 2 kb upstream and downstream of the gene body. <sup>d</sup> Genes with amino acid changes, including missense mutations in the CDS region and with 3n InDels in the CDS region. <sup>e</sup> Genes with large-effect mutations, including 3n ± 1 InDels in the CDS region, start-codon mutations, stop-codon mutations, splice-acceptor mutations, and splice-donor mutations. <sup>f</sup> Genes with large structural variations, including at least one CDS missing or other structural variation. <sup>g</sup> The total number of genes included for the analysis (genes and their orthologs in the counterpart genome, anchored on the 13 chromosomes)

### Gene-order and structural variation between *G. thurberi* and *G. davidsonii*

To analyze gene order, a total of 32,981 orthologous pairs were identified between *G. thurberi* and *G. davidsonii*, among which 1104 orthologous gene pairs were located in the inversion regions (Table 2); these account for ~3.3% of the total analyzed orthologous gene pairs. The fact that this represents a higher proportion than that between the *G. hirsutum* cultivars TM-1 and ZM24 supports that more genes are affected by interspecies inversions compared to intraspecies inversions.

We identified 21,262 (~67%) orthologous gene pairs with only missense mutations in their CDS or non-frameshift InDels between *G. thurberi* and *G. davidsonii*. However, only ~9% of the orthologous gene pairs had no amino acid changes between *G. thurberi* and *G. davidsonii*, and approximately 2% and 1% of these pairs had no variation in coding sequence (CDS) or gene bodies (CDS and intron regions), respectively (Table 2). These proportions are significantly lower than those from the comparison of *G. hirsutum* cultivars TM-1 and ZM24 (71%, 69%, and 56%), indicating that interspecies orthologs are more divergent than intraspecies homologs. Note that more than 9% of the syntenic orthologous genes pairs carried large-effect mutations, including  $3n \pm 1$  InDel, start-codon mutation, stop-codon mutation, splice-acceptor mutation, and splice-donor mutation in the CDS regions (Table 2). More than 11% of the syntenic orthologous gene pairs were impacted by large structural variations, with 85% of these having lost at least one exon; any biological significance of these variations will require further study.

We also characterized extent of gene amplification in the *G. thurberi* and *G. davidsonii* genomes. More than 3400 tandem duplicated genes were identified in both *G. thurberi* and *G. davidsonii*, among which the stress-related pathways phenylalanine metabolism, glutathione metabolism, plant-pathogen interaction, and phenylpropanoid biosynthesis were found to be enriched in a KEGG analysis, indicating that tandem duplication has apparently enhanced the tolerance of *G. thurberi* and *G. davidsonii* to various stresses (Additional file 2: Fig. S20). In total, 3136 and 3154 genes were identified as singleton genes in *G. thurberi* and *G. davidsonii*, respectively (Additional file 1: Table S9). It was notable that there was a much higher proportion of transcription factors in the whole-genome duplication and segmental duplication sets than those from singleton genes (Fisher's exact test,  $P < 2.2e-16$ ), supporting our previous finding [4] that transcription factors have a tendency to be retained after duplication (Additional file 1: Table S10).

### Identification of centromeres using a Hi-C heatmap method

Centromeres are mainly composed of repetitive retrotransposons and satellite repeats, and the challenge of accurately assembling centromeres using short-read sequencing data is well-documented [28]; accordingly, centromere evolution is poorly understood. Previous studies of Hi-C matrices have shown that centromeres form a unique type of interacting subcompartment which can function as a barrier and prevent intra-chromosomal arm interactions [29]. By exploiting the insulation feature of centromeres in Hi-C heatmap data, we successfully developed a new method for centromere characterization based on Hi-C data.

In this method, we first map the Hi-C contact data against its corresponding reference genome to obtain valid read pairs (Fig. 5a). Next, we use the valid read pairs to generate a Hi-C heatmap (at 50 kb resolution), and then use this to search regions which apparently form barriers to intra-chromosomal arm interactions. Testing confirmed that these regions, which have less frequent contacts between chromosome arms on either side compared with their frequency of intra-arm contact, are indeed centromeres (Fig. 5b). Thirdly, based on the phylogenetic relationship, we used the known cotton centromeric LTRs to align against the reference genomes to validate these Hi-C centromeres (Fig. 5c). Finally, the centromere sequence features—including sequence composition, LTR insertion time, LTRs insertion pattern, and centromeric enriched LTRs—can be cataloged systematically to support studies of centromere evolution (Fig. 5d,e). Using this new method (Additional file 2: Fig. S21), we successfully identified the centromeres in the model plant *Arabidopsis thaliana*, *Oryza sativa*, and the new *G. thurberi* and *G. davidsonii* assemblies (Additional file 2: Fig. S22-S25).

As we used nanopore long reads for our new genome assemblies, the centromeres are well assembled with the excellent coverage (Additional file 2: Fig. S26), thereby providing an unprecedented opportunity to study cotton centromere evolution. As we aligned *G. thurberi* against the *G. davidsonii* genome, we clearly found that there were no collinearities in the middle region of each orthologous chromosome (Additional file 2: Fig. S27). Chromosomal collinearity analysis showed that many non-syntenic regions were located in the centromeric regions (Additional file 1: Table S11), indicating that the centromeric regions have higher divergence compared to their neighboring (flanking) regions.

To further support this, we aligned the previously reported *G. raimondii* and *G. hirsutum* CENH3 ChIP-Seq data against the four genomes available for D genome species (*G. thurberi*, *G. davidsonii*, and Gh\_D<sub>11</sub>) (Additional file 2: Fig. S28). We detected a strong peak in a

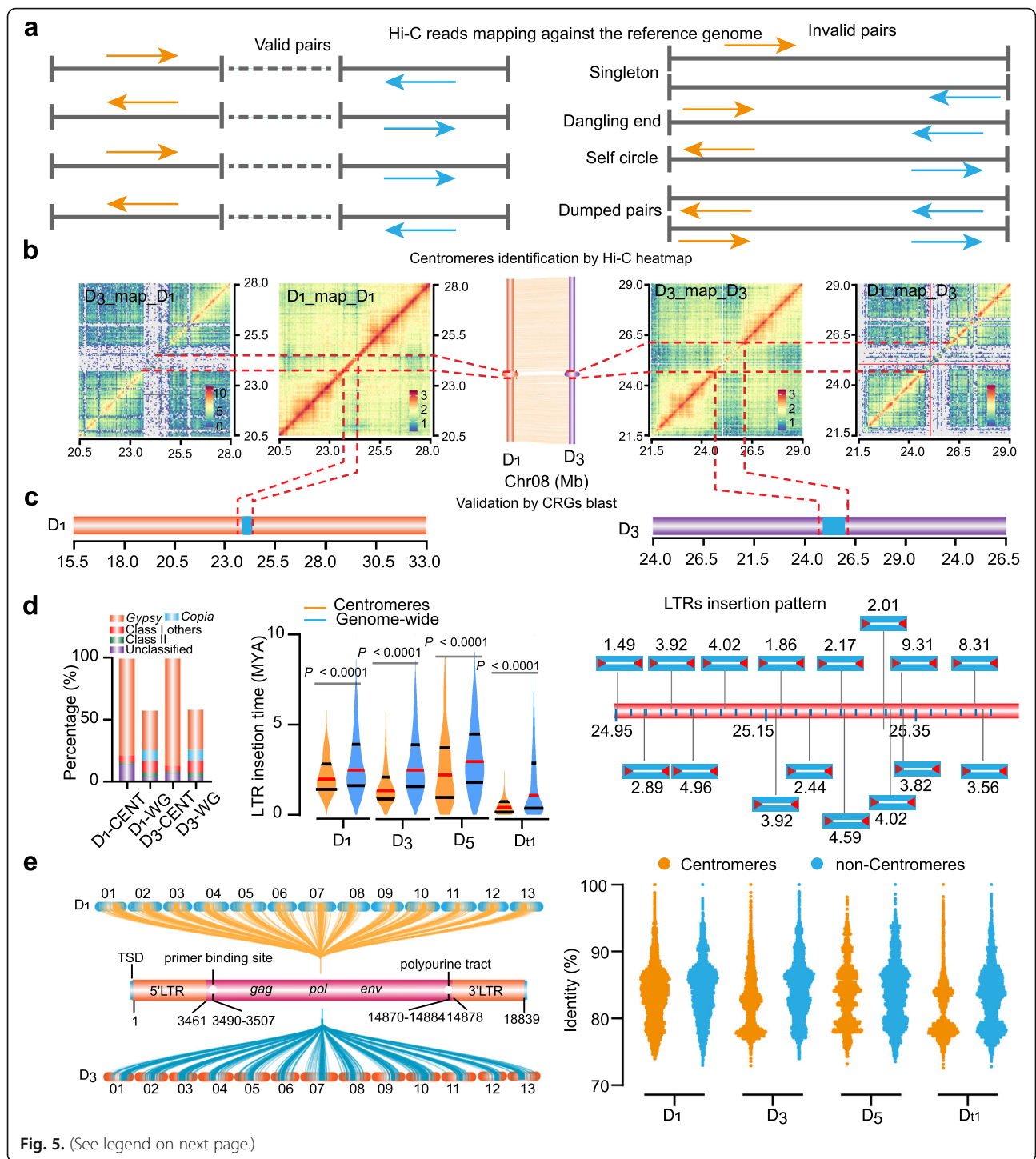


Fig. 5. (See legend on next page.)

narrow region on *G. raimondii* Chr8 when mapping *G. raimondii* ChIP-Seq data (Additional file 2: Fig. S28a). However, upon mapping *G. raimondii* ChIP-Seq data against the other three examined *Gossypium* genomes (*G. thurberi*, *G. davidsonii*, and *G. raimondii*), the signals were dispersed over a broader region, with no obvious major peaks. Mapping *G. hirsutum*

CENH3 ChIP-Seq data against the *G. hirsutum* genome revealed an apparent peak on D12; no major peaks were detected when we mapped this data to the four other D genomes (Additional file 2: Fig. S28b). These findings underscore that centromeric regions can be highly divergent among closely related species.

(See figure on previous page.)

**Fig. 5.** An overview of centromere identification based on Hi-C data. **a** A diagram of Hi-C data mapping against the reference genome. **b** Characterization of centromeres in Hi-C heat maps. The left panel shows chromatin interactions, including *G. davidsonii* mapped to *G. thurberi* ( $D_3\_map\_D_1$ ) and *G. thurberi* mapped to *G. thurberi* ( $D_1\_map\_D_1$ ). The middle panel presents a genomic alignment around the centromeres. The three-dimensional rings indicate the centromeres. The right panel shows chromatin interactions, including *G. davidsonii* mapped to *G. davidsonii* ( $D_3\_map\_D_3$ ) and *G. thurberi* mapped to *G. davidsonii* ( $D_1\_map\_D_3$ ). The regions within the orange lines are the centromere regions. **c** Validation the centromeres by centromeric LTR (Centromere Retroelement *Gossypium*, CRG) BLAST analysis. The data showed the validation on Chr08. **d** Centromere feature analysis. The right panel presents a comparison of the repetitive elements for centromeres vs. the whole genome. The middle shows LTR insertion time distributions for centromeres specifically, and for the whole genome. The center red line in the plot indicates the median, and the black lines indicate the upper and lower quartiles for insertion times. The right panel shows an analysis of the intact LTR insertion pattern. An example is presented for *G. thurberi* Chr04. The digits present the insertion time of nearby LTRs. **e** Analysis of centromere LTR enrichment. The left panel represents the sequence identity characteristic of a “CentLTR” sequence, as examined in centromeres and non-centromeric regions in four D genomes. The right panel is the identity distribution pattern of CenLTR hits presented as a dot plot. This analysis detected a total of 152,285 CenLTRs in  $D_1$  centromeres, with 163,217 in  $D_1$  non-centromeric regions; 158,815 in  $D_3$  centromeres, with 139,231 in  $D_3$  non-centromeric regions; 16,093 in  $D_5$  centromeres, with 76,875 in  $D_5$  non-centromeric regions; and 80,537 in Gh\_ $D_{t1}$  centromeres, with 246,791 in Gh\_ $D_{t1}$  non-centromeric regions

We also mapped the *G. thurberi* Hi-C data against the *G. davidsonii* assembly and vice versa, we observed large gaps in the centromeric regions; this indicates that centromeric sequences from the orthologous chromosomes in *G. thurberi* and *G. davidsonii* were highly divergent (Fig. 5b and Additional file 2: Fig. S3-4). Although the centromeric regions are highly divergent (without any syntenic blocks), we found that the flanking regions of the centromeres are highly conserved with good collinearities. For Chr03, Chr04, Chr07, and Chr08, no large-scale inversions were detected between orthologous chromosomes, highlighting that chromosomes arms are highly syntenic and lack obvious changes in their centromeric positions. Although there were inversions located in the chromosome arms in Chr05, Chr06, and Chr10, we observed that these inversions had no effect on centromere locations, since the centromeric flanking regions retained synteny. Chr01, Chr02, Chr09, Chr11, Chr12, and Chr13 experienced pericentromeric inversions; that is, we observed that the collinearities of flanking regions were reversed between the two genomes, suggesting that inversions spanning the centromere occurred after divergence.

### Centromere LTRs have undergone rapid changes

We next examined whether there were any local sequence similarities among the centromeres from non-homologous chromosomes. We used the NCBI blastn tool to align the centromere sequences, and filtered the results with a loose filter (block length larger than 2000 bp with 95% identity). We observed that the centromeric sequences are highly repetitive, and detected more similar sequences from the intraspecies comparison than the interspecies comparison, indicating that centromeres have experienced duplication after speciation (Additional file 2: Fig. S29). Moreover, we found that the sequences from *G. davidsonii* are more similar, indicating that the duplications occurred later than those from *G. thurberi*.

The DNA sequences of plant centromeres usually contain many copies of simple tandem repeats, which occur in head-to-tail arrays; only those which are associated with CENH3 nucleosomes are considered to be part of the functional centromere [30]. However, our understanding of the role of these sequences in centromere function remains rudimentary at best. Unlike centromere tandem repeats in many plants [31], we found that the tandem repeat content is very low in *G. thurberi* and *G. davidsonii* (Fig. 5d). Instead, we observed strong enrichment for LTRs (especially for *Gypsy*-type retrotransposons), suggesting that cotton centromeres have arisen from retrotransposons.

We used Kimura to analyze LTR insertion times, which revealed that LTRs in centromeres are younger than those at the whole-genome level among all D genomes ( $D_1$ ,  $D_3$ ,  $D_5$  and Gh\_ $D_{t1}$ ) (Fig. 5d). The LTRs in *G. davidsonii* centromeres are younger than those from *G. thurberi* (median of 1.336 MYA vs. 1.979 MYA), indicating that centromeres in *G. davidsonii* have been much more active than those of *G. thurberi* and supporting that the centromeres in *G. davidsonii* experienced expansion compared with those from *G. thurberi* (Fig. 5d). Unlike the nested insertion of full-length LTRs previously reported for *Brassica nigra* and some cereal centromeric regions [20], we detected full-length LTRs that were independently inserted into the centromeric region, e.g., in Chr04 of *G. thurberi*, and we identified 16 intact *Gypsy*-type LTRs that have inserted into centromeres between 1.49 and 9.31 MYA (Fig. 5d).

We constructed a phylogenetic tree of all the LTRs to describe the pattern of diversity (Additional file 2: Fig. S30). Three subclades were mainly found in the centromeric region; these were all quite distinct in sequence from the D cotton genome LTRs from non-centromere regions (Additional file 2: Fig. S30a). Moreover, we found that the LTRs from *G. davidsonii* tend to cluster together in the phylogenetic tree, as did those from *G. thurberi*, findings which indicate that the LTRs of the

centromeres in *G. thurberi* and *G. davidsonii* have proliferated and spread after these two species diverged from their common ancestor (Additional file 2: Fig. S30b).

We next identify and characterize the centromeric LTRs by mapping all the intact LTRs in the *G. thurberi* and *G. davidsonii* genomes with blastn. One LTR from Chr12 (26,780,294–26,783,754) of *G. thurberi* had significant BLAST hits for centromeres of each orthologous chromosome in *G. thurberi* and *G. davidsonii* (Fig. 5e), and we detected a variety of highly similar sequences throughout the centromeres (this LTR type was designated as “CenLTR”). Further, alignments clearly indicated strong divergence from centromere LTR types (GhCR1-GhCR4) from *G. hirsutum* (Additional file 1: Table S12).

We further aligned the *G. raimondii* and Gh\_D<sub>11</sub> genomes and found that the CenLTRs are also enriched in the centromeric their regions, indicating that CenLTRs are apparently widely distributed in the centromeres of D genome species. We compared the sequence identities between the centromeres and the non-centromere sequences for each species. A lot of CenLTR polymorphisms were detected between *G. davidsonii* centromeres and *G. davidsonii* non-centromere sequences (Fig. 5e). Similar CenLTR polymorphisms were evident between Gh\_D<sub>11</sub> centromeres and non-centromere sequences (Fig. 5e). Surprisingly, the identity with consensus sequence was lower in the centromeric regions compared with non-centromeric regions (Fig. 5e), indicating that the LTRs have undergone rapid changes in the centromeres.

#### Divergent evolution of genes involved in stress tolerance

As the D subgenome donor of the widely cultivated upland cotton, *G. raimondii* is known to have contributed stress tolerance traits to allotetraploid cotton [32]. Nevertheless, allotetraploid cotton is sensitive to *Verticillium dahliae* infection and to growth in high salinity soils; these represent major challenges facing cotton production worldwide, and a lack genetic resources for improving plant tolerance to these challenges is a major constraint in current cotton breeding programs. Here, we found that *G. thurberi* seedlings are more tolerant to *Verticillium dahliae* than *G. raimondii*, indicating that *G. thurberi* is a promising resource for upland cotton improvement (Fig. 6a). We identified 3472 and 5042 genes associated to tolerance to *Verticillium dahliae* in *G. thurberi* and *G. raimondii*, respectively. We identified a total of 106 genes including *NB-LRR*, *NPR1/3/4*, *TGA*, and downstream transcriptional factors (e.g., *WRKY33*, *SARD1*, and *CPB60g*) potentially involved in disease responses based on their differential responses to the *Verticillium dahliae* treatments between *G. thurberi* and

*G. davidsonii* (Fig. 6b). The SA biosynthesis signal pathway was activated in *G. thurberi*, as the *PAD4*, *EDS1*, *SAMT*, and *SBPB2* genes were upregulated in *G. thurberi* upon *Verticillium dahliae* challenge (Fig. 6c). We overexpressed *WRKY33* (*Gthurberi12G176500*) genes in *G. hirsutum* to test whether the genes from wild cotton can be used in cultivated cotton improvement. As expected, the overexpression lines displayed improved upland cotton tolerance to *Verticillium dahliae*, indicating that *G. thurberi* can be understood as an important genetic resource for cotton breeding (Additional file 2: Fig. S31).

Unlike *G. thurberi*, *G. davidsonii* displayed significant salt tolerance in seedlings when compared with *G. raimondii* (Fig. 6d). A total of 14 ethylene-related genes (including *SAM*, *ACS*, *ACO*, *EIN4*, *CTR*, and *EIN3*) showed differential responses to salt treatment between *G. davidsonii* and *G. raimondii* (Fig. 6e). Genes of the *CBL-CIPK* pathway showed differential responses to salt between *G. davidsonii* and *G. raimondii*, with the *CIPK* and *NHX* genes being upregulated by salt treatment of *G. davidsonii* (Fig. 6e). Moreover, we found that other well-known stress-related genes including *ERFs*, *GRASs*, *WRKY*, *NACs*, and *MYBs* were upregulated in *G. davidsonii* upon salt treatment (Fig. 6f); such genes have likely played important roles in species divergence and have likely contributed to the spread of the cotton D genome sister species in their adaptation to new ecological contexts and environments.

#### Discussion

The most outstanding advantage of long-read sequencing is that it provides more comprehensive coverage for the genome, which is often most obviously reflected by its increased capacity to accurately capture highly repetitive sequences. There was 39.4 to 46.9% of genome assembly identified as repetitive elements in the diploid draft D genomes by short-read sequencing [7], here the annotated repetitive DNA increased to approximate 60% of total assemblies.

Centromeres are known to comprise highly repetitive elements that are structures essential for the maintenance of karyotype integrity during meiosis, ensuring the fertility of developed gametes through strict inheritance of full chromosome complements [33]; nevertheless, centromeres are enigmas in many genome assemblies. Centromeres exhibit profound complexity; their length ranges from only hundreds of base pairs to multi-megabases. The simplest “point” centromeres are only ~ 125 bp and found in yeasts, and their sequences are conserved among sister species. In contrast, “regional centromeres” are significantly variable both in size and sequence and can be hundreds of kilobases in length [34]. Some regional centromeres contain non-repetitive

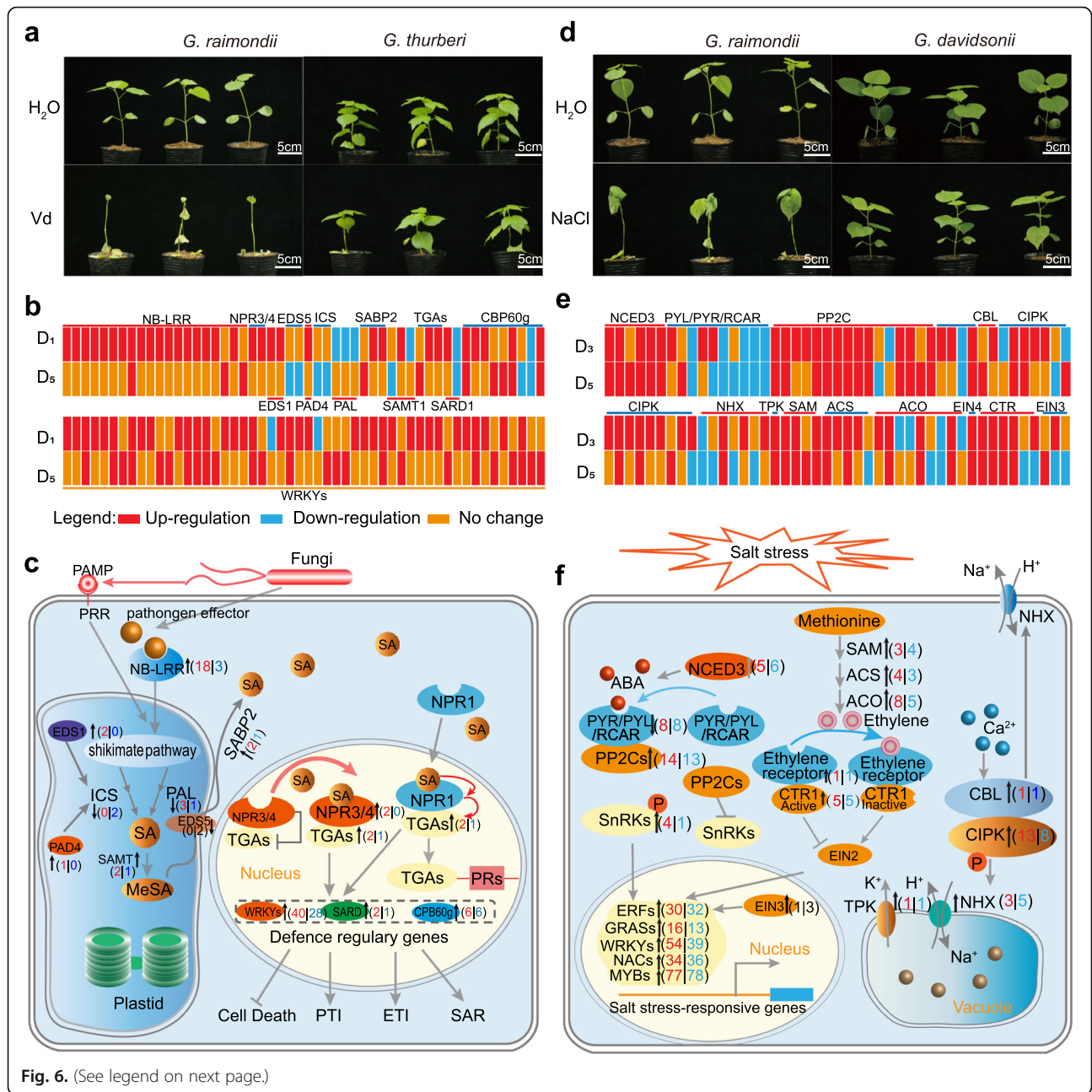


Fig. 6. (See legend on next page.)

(See figure on previous page.)

**Fig. 6.** Models depicting the molecular basis of *Verticillium* wilt and salt stress tolerance in *G. thurberi* and *G. davidsonii*. **a** Phenotypic comparison of *G. thurberi* (D<sub>1</sub>) and *G. raimondii* (D<sub>5</sub>) seedlings (35-day-old seedlings) in response to challenge with *Verticillium dahliae*. Photographs were taken under normal conditions or 14 days after challenge with *Verticillium dahliae*. **b** Heat maps for differentially expressed genes with annotations related to salicylic acid (SA) signaling, NB-LRR, and WRKYs. Genes with an adjusted *P* value < 0.05 and an absolute value of  $\log_2[\text{foldchange}] > 1$  found by EdgeR were designated as differentially expressed. **c** A proposed model showing that the SA signaling pathways enhance *Verticillium* wilt tolerance in *G. thurberi*. *V. dahliae* attack induces SA biosynthesis via the isochorismate synthase (ICS) and phenylalanine ammonia-lyase (PAL) pathways in plastids. Enhanced disease susceptibility (*EDS1*) and phytoalexin deficient 4 (*PAD4*) are required for increased SA accumulation. SA methyltransferase (*SAMT*) catalyzes SA to MeSA, which diffuses into the cytoplasm, where it is converted back to active SA by *SABP2*. The red and blue digits in brackets represent the upregulated genes in D<sub>1</sub> and D<sub>5</sub>, respectively. **d** Phenotypic comparison of *G. davidsonii* (D<sub>3</sub>) and *G. thurberi* (D<sub>1</sub>) seedlings in response to salt stress treatment (250 mM NaCl watering 21-day-old seedlings every 2 days). Photographs were taken under normal conditions or 14 days after treatment with NaCl solution. **e** Heat maps for differentially expressed genes with annotations related to ABA, ethylene, and CBL-CIPK pathways. Genes with an adjusted *P* value < 0.05 and an absolute value of  $\log_2[\text{foldchange}] > 1$  found by EdgeR were designated as differentially expressed. **f** Transcriptional network related to salt response in *G. raimondii* and *G. davidsonii*. Ethylene biosynthesis, calcium signaling, and vacuole NHX are activated in *G. davidsonii*. The *NCED3* gene encodes the enzyme which catalyzes the first step of ABA biosynthesis. The ABA signaling pathway, comprising *PYR/PYL/RCAR*, *PP2C*, and *SnRKs* proteins, is a major plant hormone involved in salt stress responses. Ethylene biosynthesis is catalyzed by the *SAM*, *ACS* (ACC synthase), and *ACO* (ACC oxidase) enzymes. The ethylene signaling pathway includes ethylene receptor, *CTR1*, and *EIN2*. *TPK* (two-pore potassium) is K<sup>+</sup> channel that trafficks K<sup>+</sup> out of the vacuole. *NHX1* (tonoplast-based Na<sup>+</sup>/H<sup>+</sup> exchanger) is required for sequestration of excessive Na<sup>+</sup> and Cl<sup>-</sup> in the vacuole. The red and blue digits in the brackets represent the upregulated genes in D<sub>3</sub> and D<sub>5</sub>, respectively

sequences, e.g., *Candida albicans*, or have a mixture of repetitive sequence and non-repetitive sequence (e.g., chicken and horse) [31]. A previous study showed that Ty3-gypsy-like LTRs are localized to the centromeric region of all the chromosomes of upland cotton and the B-, D-, and E-genome diploid cottons [35]. In the present study, we found these large numbers of gypsy-like LTR in the centromeric regions account for more than 75% of the total centromere length, strongly supporting that the cotton centromeres originated from retrotransposons.

The precise centromeric DNA sequences of the *G. thurberi* and *G. davidsonii* vary dramatically, and it has been proposed that this rapid evolution could be a consequence of meiotic drive [13, 16]. Comparison between the closely related species human and macaque showed that some centromeres adopt new positions over evolutionary time subsequent to a speciation event, without transposing any surrounding genetic markers. These structures are referred to as evolutionarily new centromeres and have been observed in primates and other mammals [13]. Unlike the human and macaque, for cotton (except for the Chr11) the centromere location is highly conserved between the orthologous chromosomes in *G. thurberi* and *G. davidsonii*.

Hi-C sequence data has recently supported a new era of studies about genome-wide 3D genome structural organization. Previous work in cotton examined how chromatin architecture reorganization may have been affected during polyploidization, in analyses based on comparisons of allotetraploid cotton with the possible progenitor of diploid cotton [24]. In the present study, we compared the 3D genome of three sister D genome species, and our observations of both A/B compartment switching and the reorganization of TADs among the

diploid D genomes provide new insights into the effects of genome divergence on the spatial organization of chromatin. Specifically, we found that the DNA sequence surrounding the TAD boundaries displayed a lower level of 5mC (CG, GHG, and CHH) and 6mA modifications than the genome as a whole; these epigenetic modifications have been functionally associated with gene transcription in previous studies and may play roles in the activation of genes positioned near TAD boundary genes.

## Conclusions

In summary, we de novo assembled very high-quality reference genomes for two important cotton germplasm resources using technologically complementary sequencing technologies. Based on these reference-grade genome assemblies for *G. thurberi* and *G. davidsonii*, we comprehensively evaluated genome arrangements, small variations (SNPs, InDels, and PAVs), gene order, long-range interactions between proximal and distal regulatory regions, and gene structure variations in these closely related species to better understand their divergence process. Of particular note, our Nanopore long-read data, in combination with a new method we developed for centromere characterization based on Hi-C data, ultimately revealed insights about the previously mysterious process of centromere evolution in plant genome. Our study also identified multiple genetic networks underlying the unique traits that have long made these D genome cotton species interesting to crop scientists. Thus, our work deepens understanding of crop evolution, centromere divergence, and trait diversification and indicates a way forward for harnessing genes that confer agronomically beneficial traits as useful resources in cotton breeding programs.



## Methods

### Plant growth conditions

*G. thurberi* and *G. davidsonii* plants (collection from National Wild Cotton Nursery, Sanya, China) were grown in a greenhouse for 180 days, and the young leaves from a single plant were harvested and immediately frozen in liquid nitrogen to extract their genomic DNA. For *Verticillium* wilt analysis, *G. raimondii* and *G. thurberi* seedlings were planted in a mixture of sand and vermiculite. Once they had two true leaves, they were treated with *Verticillium dahliae* via the root dipping method according to the methods used in a previous study [36]. For salt tolerance analysis, *G. raimondii* and *G. davidsonii* were planted in vermiculite and watered with 350 mmol of NaCl every 2 days. All of the phenotypic photos were taken 14 days after treatment. For mRNA-seq analysis, the seedlings were hydroponically cultivated according to previously used methods [36]. Seedlings at the two-true-leaf stage were treated with a liquid medium containing 250 mmol of NaCl; seedlings grown in a normal liquid medium were used as the control. The samples were harvested at 3 h, 6 h, and 24 h following treatment, after which they were immediately frozen in liquid nitrogen and stored at  $-80^{\circ}\text{C}$  for subsequent analysis.

### Nanopore sequencing

An improved CTAB method [6] was used to extract the genomic DNA of *G. thurberi* and *G. davidsonii*. Two micrograms of gDNA was repaired using a NEB Next FFPE DNA Repair Mix kit (M6630, USA) and subsequently processed using the ONT Template prep kit (SQK-LSK109, UK) according to the manufacturer's instructions. Large segment libraries were premixed with loading beads and washed with an R9 flow cell. The library was sequenced on the ONT PromethION platform with a corresponding R9 cell and ONT sequencing reagent kit (EXP-FLP001.PRO.6, UK) according to the manufacturer's instructions.

MinKNOW software was used to collect the sequencing data in real-time and process it into basecalls. Single-molecule sequencing was performed on a PromethION system and yielded a total of 3,575,506 and 3,237,739 filtered subreads with average lengths of 31,963 bp and 33,450 bp for *G. thurberi* and *G. davidsonii*, respectively. Finally, only nanopore subreads equal to or longer than 500 bp were used to generate the two genome assemblies.

Total RNA extraction and cDNA synthesis were performed according to the methods described by Yang et al. [4]. The BluePippin™ Size Selection System (Sage Science, USA) was used to identify and select the requisite sizes (1–2 kb, 2–3 kb, and > 3 kb), and a Pacific Biosciences DNA Template Prep Kits v.2.0 was used to build the SMRT bell libraries. We conducted the SMRT

sequencing using the Pacific Bioscience Sequel platform according to the manufacturer's instructions.

### Illumina sequencing

We constructed libraries with a 350-bp insert fragment for *G. thurberi* and *G. davidsonii* according to the manufacturer's instructions (Illumina). A HiSeq 2500 system was used to sequence the libraries, along with a PE150 strategy according to the manufacturer's instructions (Illumina). The sequence adaptors were removed for the uncleaned Illumina reads, and the contaminated reads (viral, mitochondrial, bacterial sequences) were compared with the NCBI-NR database via BWA v0.7.13 [37] (using default instructions). Duplicate pairs were identified using FastUniq v1.12 [38]. In total, we produced 116.9 Gb and 118.2 Gb clean Illumina reads for *G. raimondii* and *G. davidsonii*, respectively. For CENH3 analysis, the raw sequencing data were downloaded from the Gene Expression Omnibus for *G. hirsutum* (accession number GSE119184) [39] and the European Molecular Biology Laboratory-European Bioinformatics Institute (accession number PRJEB14368) [40]. The data were aligned to the reference genome with Botiwe2, and the enrichment was calculated by dividing the CENH3 read counts by the input read counts according to previously used methods [28].

### De novo assembly

Canu [41] (<https://github.com/marbl/canu>, v1.5) was used to select longer seed reads with the settings “genomeSize = 1000000000” and “corOutCoverage = 50,” after which raw overlapping reads were detected using a highly sensitive overlapper MHAP (mhap-2.1.2, option “corMhapSensitivity = low/normal/high”). An error correction was then performed using falcon\_sense method (option “correctedErrorRate = 0.025”). Smartdenovo (<https://github.com/ruanjue/smartdenovo>) was used for assembly. Racon software was used for error correction, and Pilon software was used for adjustment [42].

The two assemblies were evaluated by mapping 1440 Benchmarking Universal Single-Copy Orthologs to the genomes using BUSCO v3.0.2 b [17], which showed 1372 (95.28%) and 1374 (98.42%) complete BUSCOs and 18 (1.25%) and 14 (0.97%) fragmented BUSCOs in the *G. thurberi* and *G. davidsonii* assemblies, respectively.

### Hi-C sequencing data

The Hi-C libraries construction and sequencing were performed according to the methods described by Yang et al. [4]. After the reads were filtered, we obtained 284.3 million and 280.3 million valid interaction pairs for the chromosome-level assembly of *G. thurberi* and *G. davidsonii*, respectively. The assembly contigs were separated into 50-kb fragments, and LACHESIS software [43] was

used to cluster those that remained with valid interaction read pairs; finally, 74 and 104 contigs with respective total lengths of 779.6 Mb and 801.2 Gb were anchored and oriented to their 13 chromosome-level groups of *G. thurberi* and *G. davidsonii*, respectively.

### Repeated sequence prediction

The repeated sequences were identified according to the methods described by Yang et al. [4]. The repeated sequences occupy 57.96% (451.8 Mb) of the *G. thurberi* assembly and 58.58% (469.4 Mb) of *G. davidsonii* assembly, respectively, of which *Gypsy* retrotransposons account for more than 31% in both assemblies. The insert time was calculated using the solo and intact LTRs according to the following:  $\text{time} = K/2r$  ( $K$  is the distance between all of the alignment pairs;  $r$  is the rate of nucleotide substitution). The  $r$  value<sup>9</sup> was considered to be  $7 \times 10^{-9}$ , while we used the distmat program to calculate  $K$ . All of this was performed using the EMBOSS<sup>30</sup> package according to the Kimura two-parameter model.

### Protein-coding gene prediction

We used Iso-Seq, protein homology, and de novo methods. De novo gene prediction entailed using GenScan v1.0 [44], Augustus v2.4 [45], GlimmerHMM v3.0.4 [46], GeneID v1.4 [47], and SNAP [48]; the homologous peptides were aligned to the assemblies from *Oryza sativa* L. ssp. japonica, *Arabidopsis thaliana*, *G. raimondii* (JGI), and *G. hirsutum* (CRI) using GeMoMa v1.4.2 [49], and BLAT [50] was used to align the consensus isoforms derived from PacBio long cDNA reads to the repeat-masked assemblies. Lastly, PASA [51] was used to analyze the gene structures of the results of the BLAT alignment. We used EVIDENCEModeler [52] to combine the protein alignments, transcript information, and de novo predictions to produce a unifying model for the gene. In total, 41,316 and 41,471 genes were predicted for *G. thurberi* and *G. davidsonii*, respectively, whereas 37,533 and 38,755 genes were annotated in the previously reported draft genomes for *G. thurberi* and *G. davidsonii* [7], respectively. We annotated the expected genes by comparing their sequences with several protein sequence and nucleotide repositories, including NR, COG, KEGG, and TrEMBL, using an  $e$ -value cutoff of  $1e^{-5}$ . Blast2GO was used to designate gene ontology (GO) terms for all genes based on NCBI databases. BUSCO v3.0.2 was compared with embryophyta\_odb10 database to ensure the gene set was complete, compared to the reference genome sequences. More than 97.8% of the BUSCO genes were complete and only 0.2% of BUSCOs were missing. This indicates that our gene prediction is accurate.

### Paralog analysis for *G. thurberi* and *G. davidsonii*

The all-against-all BLASTP method ( $e$ -value  $<1e^{-5}$ ) was used to detect paralogous genes in *G. thurberi* and *G. davidsonii*. Homologous blocks were then detected using MCScanX v1.1 [53], requiring at least five collinear gene pairs within one block and fewer than 25 intervening genes.

### Identification of SNPs, InDel, inversions, and PAVs

We used MUMmer v4.0 [54] (<http://mummer.sourceforge.net/>) to identify the SNPs and InDel between *G. thurberi* and *G. davidsonii* according to the following procedures: (1) each query genome was aligned with the corresponding reference genome using the nucmer utility under the parameter “-mum,” (2) the delta-filter utility was used to filter mapping noise and determine the one-to-one alignment blocks with the parameters “-1 -r -q,” and (3) the show-snps utility was used for calling the SNP and small InDels ( $<100$  bp).

Inversions were obtained by screening nucmer outputs with a delta-filter utility using two parameters: “-i 90 -g -r -q” and “-i 90 -1 -r -q,” in which “-1” identified one-to-one alignment blocks for rearrangements, and “-g” identified collinear regions with global alignments and no rearrangements. Alignment blocks with translocated sections associated with global alignments under the -g parameter were considered allelic and were not included. We identified non-allelic parts by comparing the allelic locations from the alignment blocks generated by the “-1” parameter, which were considered translocations or inversions based on their positioning in the overall region. Delta filtering was used to screen nucmer output, with a minimum identity of 90%. Homology blocks not associated with the allelic locations were considered translocations or inversions based on their positioning in the overall region. The PAV of *G. thurberi* or *G. davidsonii* were called by ppsPCP ( --coverage 0.5 --sim\_pav 0.9).

### Genes and their structural variations analysis

The one-to-one orthologous genes among *G. arboreum*, *G. thurberi*, *G. davidsonii*, *G. raimondii*, and *G. turneri* were identified using Inparanoid v4.1 [55]. In some instances, no matching counterpart was detected in the subject genome for an ortholog in the query genome. In these cases, the GeMoMa software was used to verify its absence in order to avoid a prediction error. In total, we obtained 31,319, 32,981, 26,237, 25,925, 31,130, and 26,057 orthologous gene pairs for *G. davidsonii*-*G. raimondii*, *G. davidsonii*-*G. thurberi*, *G. davidsonii*-*G. turneri*, *G. raimondii*-*G. turneri*, *G. thurberi*-*G. raimondii*, and *G. thurberi*-*G. turneri*, respectively. Similarly, we identified 28,704, 28,495, 28,631, and 25,063 orthologous pairs between *G. arboreum* and *G. davidsonii*, *G. raimondii*,

*G. thurberi*, and *G. turneri*, respectively. The CLUSTALW was used to align the coding and protein sequences of the orthologous pair, while the *Ka* and *Ks* values were calculated using the perl module, “Bio::Align::DNAStatistics.” Divergence time estimation was followed a recent study in allotetraploid cotton [56].

We assessed the genetic variations within each orthologous gene pair according to the following parameters: (1) we identified the longest transcript within each gene loci as a candidate; (2) we obtained the gene coding regions 2 kb upstream and 2 kb downstream and compared them with their associated genomic regions with BWA; (3) we considered structurally conserved genes to be orthologous genes with the same number of exons but with asynchronous differences and a lack of certain codons; (4) large-effect variations were considered to be orthologous genes with the same number of exons, but varying lengths, different splice sites, or mutations in their frameshifts; (5) we considered the orthologous genes that were left to have significant structural differences.

#### Gene family expansion analysis

Protein sequences with lengths less than 20 amino acid residues were filtered. The blastp program was used to perform all-vs-all alignments with “-evalue 1e-5, -outfmt 6.” OrthomclBlastParse within OrthoMCL [57] was used to filter the blastp results to become inputs of the Mysql database. The orthmclPair was used to identify potential protein pairs, while orthmclDumpPairFiles was used to obtain the orthologous pairs. The mcl was used to cluster the output of the orthmclDumpPairFiles, while orthomclMclToGroups was used to group the orthologous pairs. The unique genes were analyzed based on grouped orthologous pairs. A total of 7561 genes were identified as single-copy genes, and their protein sequences were combined as an ultra-long fasta species by species. The sequences were then aligned via MAFFT [58], and the conserved sites were extracted by Gblocks [59]. The optimal amino substitution model evaluated by Prottest software [60] was “PROTGAMMAJTTX” for phylogenetic tree analysis using RAxML [61]. Previous studies, including this one, demonstrated that the diploid A and D genome species were divergent at ~5 MYA and that *G. thurberi* and *G. davidsonii* were divided at 1.40–1.60 MYA. These two time points were used to fix the node divergent time in the r8s [62] analysis. CAFE ([http://heanet.sourceforge.net/project/cafehahnlab/cafe.linux.x86\\_64](http://heanet.sourceforge.net/project/cafehahnlab/cafe.linux.x86_64)) was used to evaluate the expansion and contraction of the gene family. A total of 2228 families experienced expansion and contraction.

#### Identification of the centromeres

The relatively conserved 5′ LTR sequences (GhCR1-5′ LTR, GhCR2-5′ LTR, GhCR3-5′ LTR, and GhCR4-5′ LTR), which are related to centromeres, have been identified in cotton [35]. Here, we aligned these LTR sequences against the *G. thurberi* and *G. davidsonii* genomes using blastn with sequence similarity  $\geq 80\%$  and *e*-value  $\leq 1e-20$ . After filtering the alignments, we used the R t.test function to calculate the 95% confidence interval for the median, which represents the centromeric region for each chromosome. Previous studies have demonstrated that centromeres form a barrier to intra-chromosomal arm interactions, resulting in less frequent contacts between the chromosome arms on either side compared with the frequency of intra-arm contact [29]. This insulation property is visible on contact maps and can be used to identify centromeres. The Hi-C reads of the *G. thurberi* were truncated at the putative Hi-C junctions using the HindIII restriction sites, followed by alignment to the D1 reference genomes using bwa (version 0.7.10). The uniquely mapping read pairs with a mapping quality greater than 20 were kept for further analysis. Invalid read pairs, including dangling-end and self-cycle, re-ligation, dumped products, and PCR duplicates were filtered using HiC-Pro software (v2.10.0). The R ggplot2 package was used to draw the Hi-C heatmaps at different resolutions. In our new assemblies, we also observed insulation features in the Hi-C heatmaps: *G. thurberi* and *G. davidsonii* centromeres lack a strong interacting signal and create a barrier within each chromosome, resulting in the expected trend for contacts in plant genomes: less frequent contacts between the chromosome arms on either side of them compared with intra-arm contact frequencies (Additional file 2: Fig. S24-25). We used the Hi-C heatmaps to shrink the centromeric region as the centromeres displayed a significantly reduced interaction with the neighboring regions. We found that centromeric regions overlapped with the result obtained by GhCRs blast. Therefore, the centromeric regions identified by Hi-C were used in further analysis.

#### Methylation sequencing analysis

We used Basecall with ONT’s Guppy software to convert fast5 format data to the fastq format for QC analysis. The original fastq data were further filtered to remove adapters, short reads (length < 500 bp), and low-quality reads (MeanQual < 6). From this, we obtained the total data set. Sequencing depth and alignment efficiency were counted by aligning clean-read positions at the reference genome after mapping them to the reference genome with minimap2 software. The minimap2 software uses split-reads and heuristic methods to reduce false comparisons, which are suitable for

comparing long-read, high-noise reads with reference genes. We used nanopolish, based on the hidden Markov model, to detect CpG. Tombo was used to detect CHH (H = A/T/C), CHG, and 6mA sites. The bisulfite sequencing (Bs-seq) and data analysis was performed according to a previous report [24].

### mRNA-seq analysis

We used hisat2 to map the clean reads (Additional file 1: Table S13-14) to the reference genome, and Stringtie to calculate the gene expression level and read counts [63]. For differentially expressed genes, we used the DESeq2 package in R, based on the negative binomial distribution. Genes with absolute value of  $\log_2[\text{fold change}] > 1$ , and Benjamini–Hochberg-adjusted  $P < 0.05$  are considered DEGs. GO enrichment was performed using the R package “goseq,” and the package “qvalue” was used to adjust the  $P$  value. Only GO terms with adjusted  $P$  value  $< 0.05$  were considered significant.

### Identification of A/B compartment, TAD boundaries, and long-range interactions

The clean reads were mapped against their corresponding reference genomes with the bowtie2 software (version 2.2.3). Based on a Hi-C Pro pipeline, only the uniquely mapped paired-end reads were used for subsequent analysis [64]. The contact matrices were generated at different resolutions using the robust remove multiplex bias method ICE. The A/B compartments were analyzed based on 50 kb resolution ICE matrices using HiTC (version 1.24.0). The eigenvalue of the first principle components was plotted as the compartment assignment, with positive values corresponding to high gene density (compartment A) and negative values corresponding to low gene density (compartment B).

TAD calling was performed by TadLib [65] with a 20-kb resolution. The inclusion ratio (IR) for each TAD was calculated by HOMER [66], with only IR values greater than 1 used in subsequent analysis. TAD boundaries were identified from the calling results of the above TAD. We compared the different directionality index (DI) delta scores of the TAD boundaries between two samples to identify dynamic TAD boundaries. A TAD boundary was considered dynamic when its adjusted  $P$  value (as calculated by LIMMA (v3.30.13) [67]) was less than 0.1 and it had DI delta scores greater than 70 for one sample and smaller than 70 for the other sample with a fold change of DI delta scores larger than 2. If an adjusted  $P$  value was larger than 0.1 and the DI delta score for one sample was four times as large as the other sample, the TAD boundary was also deemed as dynamic [68]. A hierarchical clustering heatmap was drawn for all dynamic TAD boundaries based on their DI delta scores.

The Fit-Hi-C (v2.05) tool was used to identify the Hi-C interaction peaks with “-r 10,000, -U 1000000, -L 20000, -x intraOnly.” Then, the  $q$ -value  $\leq 0.00001$  was used to filter the candidate interaction sites. HOMER was used to calculate the differential chromosomal interactions. Hi-C peaks 2-kb upstream or 1-kb downstream of the TSS of the genes were annotated as proximal Hi-C peaks (P), and the others were annotated as distal Hi-C peaks (D).

### Statistical analysis

Comparison of methylation levels between the A and B compartments was carried out by using a two-sided Wilcoxon rank-sum test.

### Abbreviations

TAD: Topologically associating domains; *G. raimondii*: *Gossypium raimondii*; *G. hirsutum*: *Gossypium hirsutum*; *G. barbadense*: *Gossypium barbadense*; *G. tomentosum*: *Gossypium tomentosum*; *G. mustelinum*: *Gossypium mustelinum*; *G. darwinii*: *Gossypium darwinii*; Hi-C: High-throughput chromosome conformation capture; CRG: Centromere Retroelement *Gossypium*

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12915-021-01041-0>.

**Additional file 1: Tables S1-S14. Table S1.** Summary of Nanopore sequencing clean data in *G. thurberi* and *G. davidsonii*. **Table S2.** Nanopore sequencing reads length distribution in *G. thurberi* and *G. davidsonii*. **Table S3.** Genome assembly statistic information for *G. thurberi* and *G. davidsonii*. **Table S4.** Summary of Hi-C contact data mapped against the *G. thurberi* and *G. davidsonii*. **Table S5.** Comparison of repetitive elements between *G. thurberi* and *G. davidsonii*. **Table S6.** The Illumina short reads from *G. thurberi* and *G. davidsonii* mapping against *G. thurberi* and *G. davidsonii* assemblies respectively. **Table S7.** Evaluation of the *G. thurberi* and *G. davidsonii* assemblies using BUSCO database. **Table S8.** Summary of 1-to-1 blocks between *G. thurberi* and *G. raimondii*, between *G. davidsonii* and *G. raimondii*, between *G. thurberi* and *G. davidsonii* or between *G. thurberi* and *G. turneri*. **Table S9.** Summary of gene duplication type information for the *G. thurberi* and *G. davidsonii* genomes. **Table S10.** The duplication gene type of transcription factors. **Table S11.** The centromeric regions identified by Hi-C heatmap combined the centromere specific CRGs mapping. **Table S12.** Comparison of CenLTR with the reported GhCR1-GhCR4. **Table S13.** Summary of Illumina sequencing clean data in *G. thurberi* and *G. davidsonii*. **Table S14.** Summary of accession information for mRNA-seq.

**Additional file 2: Figures S1-S31.**

### Acknowledgements

Not applicable.

### Authors' contributions

F.G.L. and Z.E.Y. conceived and designed the research. Z.E.Y. and F.G.L. managed the project. F.Y.L. and Z.E.Y. performed the genome sequencing, assembly, and bioinformatics. Z.E.Y., X.Y.G., W.N.L., Y.Y.J., L.S.L., W.H., and Y.L.C. prepared the samples, performed phenotyping, and contributed to data analysis. Z.E.Y. and F.G.L. designed the molecular experiments and Y.Y.J., L.S.L., and W.H. performed the molecular experiments and led interpretation of the molecular data analysis. Z. E.Y. prepared the figures and tables. Z.E.Y., X.Y.G., S.L.P., and F.G.L. wrote and revised the manuscript. All authors read and approved the final manuscript.

### Funding

This work was supported by funding from the National Natural Science Foundation of China (grants 31621005 and 31690093 to F.G.L.), Agricultural

Science and Technology Innovation Program of Chinese Academy of Agricultural Sciences, Central Public-interest Scientific Institution Basal Research Fund (Y2020PT13) and Young Elite Scientists Sponsorship Program by CAST (2019-2021QNR0001 to Z.E.Y.).

#### Availability of data and materials

All the raw sequencing data for *G. thurberi* and *G. davidsonii* genome assembly data, mRNA-seq, and methylation are accessible through the NCBI under accession PRJNA659592. These supporting data (genome assemblies and genes annotations, as well as gff files for gene models) are available from the website ([grand.cricaas.com.cn](http://grand.cricaas.com.cn)). Data supporting the findings of this work are available within the paper and its Supplementary Information files. The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

#### Declarations

##### Ethics approval and consent to participate

Not applicable.

##### Consent for publication

Not applicable.

##### Competing interests

The authors declare that there are no competing interests.

#### Author details

<sup>1</sup>Zhengzhou Research Base, State Key Laboratory of Cotton Biology, Zhengzhou University, Zhengzhou 450001, China. <sup>2</sup>Institute of Cotton Research, Chinese Academy of Agricultural Sciences, Anyang 455000, China. <sup>3</sup>College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China. <sup>4</sup>School of Computer Science, National University of Defense Technology, Changsha 410073, China. <sup>5</sup>Peng Cheng Lab, Shenzhen 518000, China. <sup>6</sup>Biomarker Technologies Corporation, Beijing 101300, China.

Received: 24 December 2020 Accepted: 29 April 2021

Published online: 03 June 2021

#### References

- Mehboob ur R, Shaheen T, Tabbasam N, Iqbal MA, Ashraf M, Zafar Y, et al. Cotton genetic resources. A review. *Agronomy Sustainable Dev.* 2011;32:419–32.
- Yang Z, Qanmber G, Wang Z, Yang Z, Li F. *Gossypium* genomics: trends, scope, and utilization for cotton improvement. *Trends Plant Sci.* 2020;25(5):488–500. <https://doi.org/10.1016/j.tplants.2019.12.011>.
- Adams KL, Wendel JF. Polyploidy and genome evolution in plants. *Curr Opin Plant Biol.* 2005;8(2):135–41. <https://doi.org/10.1016/j.cpb.2005.01.001>.
- Yang Z, Ge X, Yang Z, Qin W, Sun G, Wang Z, et al. Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. *Nat Commun.* 2019;10(1):2989. <https://doi.org/10.1038/s41467-019-10820-x>.
- Du X, Huang G, He S, Udall J, Yoo MJ, Byers R, Chen W, Doron-Faigenboim A, Duke MV, Gong L, Grimwood J, Grover C, Grupp K, Hu G, Lee TH, Li J, Lin L, Liu T, Marler BS, Page JT, Roberts AW, Romanel E, Sanders WS, Szadkowski E, Tan X, Tang H, Xu C, Wang J, Wang Z, Zhang D, Zhang L, Ashrafi H, Bedon F, Bowers JE, Brubaker CL, Chee PW, Das S, Gingle AR, Haigler CH, Harker D, Hoffmann LV, Hovav R, Jones DC, Lemke C, Mansoor S, Rahman M, Rainville LN, Rambani A, Reddy UK, Rong JK, Saranga Y, Scheffler BE, Scheffler JA, Stelly DM, Triplett BA, van Deynze A, Vaslin MFS, Waghmare VN, Walford SA, Wright RJ, Zaki EA, Zhang T, Dennis ES, Mayer KFX, Peterson DG, Rokhsar DS, Wang X, Schmutz J: Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* 2012, 492:423–427, 7429, doi: <https://doi.org/10.1038/nature11798>.
- Zhu G, Li W, Zhang F, Guo W. RNA-seq analysis reveals alternative splicing under salt stress in cotton, *Gossypium davidsonii*. *BMC Genomics.* 2018; 19(1):73. <https://doi.org/10.1186/s12864-018-4449-8>.
- Ulloa M. The diploid d genome cottons (*Gossypium* spp.) of the new world. In: *World Cotton Germplasm Resources*; 2014.
- Udall JA, Long E, Hanson C, Yuan D, Ramaraj T, Conover JL, et al. De novo genome sequence assemblies of *Gossypium raimondii* and *Gossypium turneri*. G3 (Bethesda). 2019;9:3079–85.
- Wang K, Wang Z, Li F, Ye W, Wang J, Song G, et al. The draft genome of a diploid cotton *Gossypium raimondii*. *Nat Genet.* 2012;44(10):1098–103. <https://doi.org/10.1038/ng.2371>.
- McKinley KL, Cheeseman IM. The molecular basis for centromere identity and function. *Nat Rev Mol Cell Biol.* 2016;17(1):16–29. <https://doi.org/10.1038/nrm.2015.5>.
- Varoquaux N, Liachko I, Ay F, Burton JN, Shendure J, Dunham MJ, et al. Accurate identification of centromere locations in yeast genomes using Hi-C. *Nucleic Acids Res.* 2015;43(11):5331–9. <https://doi.org/10.1093/nar/gkv424>.
- Huang G, Wu Z, Percy RG, Bai M, Li Y, Frelichowski JE, et al. Genome sequence of *Gossypium herbaceum* and genome updates of *Gossypium arboreum* and *Gossypium hirsutum* provide insights into cotton A-genome evolution. *Nat Genet.* 2020;52(5):516–24. <https://doi.org/10.1038/s41588-020-0607-4>.
- Grover CE, Pan M, Yuan D, Arick MA, Hu G, Brase L, et al. The *Gossypium longicalyx* genome as a resource for cotton breeding and evolution. G3 (Bethesda). 2020;10:1457–67.
- Seppy M, Manni M, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness. *Methods Mol Biol.* 1962;2019:227–45.
- Zhao FA, Fang WP, Xie DY, Zhao YM, Tang ZJ, Li W, et al. Proteomic identification of differentially expressed proteins in *Gossypium thurberi* inoculated with cotton *Verticillium dahliae*. *Plant Science.* 2012;185:176–84.
- Wang M, Tu L, Yuan D, Zhu SC, Li J, Liu F, et al. Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. *Nat Genet.* 2019;51(2):224–9. <https://doi.org/10.1038/s41588-018-0282-x>.
- Perumal S, Koh CS, Jin L, Buchwaldt M, Higgins EE, Zheng C, et al. A high-contiguity *Brassica nigra* genome localizes active centromeres and defines the ancestral *Brassica* genome. *Nat Plants.* 2020;6(8):929–41. <https://doi.org/10.1038/s41477-020-0735-y>.
- Liu Q, Fang L, Yu G, Wang D, Xiao CL, Wang K. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nat Commun.* 2019;10(1):2449. <https://doi.org/10.1038/s41467-019-10168-2>.
- Ni P, Huang N, Zhang Z, Wang DP, Liang F, Miao Y, et al. DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics.* 2019;35(22):4586–95. <https://doi.org/10.1093/bioinformatics/btz276>.
- Zhang Q, Liang Z, Cui X, Ji C, Li Y, Zhang P, et al. N(6)-methyladenine DNA methylation in Japonica and Indica rice genomes and its association with gene expression, plant development, and stress responses. *Mol Plant.* 2018; 11(12):1492–508. <https://doi.org/10.1016/j.molp.2018.11.005>.
- Wang M, Wang P, Lin M, Ye Z, Li G, Tu L, et al. Evolutionary dynamics of 3D genome architecture following polyploidization in cotton. *Nat Plants.* 2018; 4(2):90–7. <https://doi.org/10.1038/s41477-017-0096-3>.
- Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, et al. Chromatin architecture reorganization during stem cell differentiation. *Nature.* 2015;518(7539):331–6. <https://doi.org/10.1038/nature14222>.
- Li E, Liu H, Huang L, Zhang X, Dong X, Song W, et al. Long-range interactions between proximal and distal regulatory regions in maize. *Nat Commun.* 2019;10(1):2633. <https://doi.org/10.1038/s41467-019-10603-4>.
- Gaudinier A, Rodriguez-Medina J, Zhang L, Olson A, Liseron-Monfils C, Bagman AM, et al. Transcriptional regulation of nitrogen-associated metabolism and growth. *Nature.* 2018;563(7730):259–64. <https://doi.org/10.1038/s41586-018-0656-3>.
- Han J, Masonbrink RE, Shan W, Song F, Zhang J, Yu W, et al. Rapid proliferation and nucleolar organizer targeting centromeric retrotransposons in cotton. *Plant J.* 2016;88(6):992–1005. <https://doi.org/10.1111/tbj.13309>.
- Muller H, Gil J Jr, Drinnenberg IA. The impact of centromeres on spatial genome architecture. *Trends Genet.* 2019;35(8):565–78. <https://doi.org/10.1016/j.tig.2019.05.003>.

30. Su H, Liu Y, Liu C, Shi Q, Huang Y, Han F. Centromere satellite repeats have undergone rapid changes in polyploid wheat subgenomes. *Plant Cell*. 2019; 31(9):2035–51. <https://doi.org/10.1105/tpc.19.00133>.
31. Comai L, Maheshwari S, Marimuthu MPA. Plant centromeres. *Curr Opin Plant Biol*. 2017;36:158–67. <https://doi.org/10.1016/j.pbi.2017.03.003>.
32. Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J, et al. Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat Biotechnol*. 2015;33(5):531–7. <https://doi.org/10.1038/nbt.3207>.
33. Barra V, Fachinetti D. The dark side of centromeres: types, causes and consequences of structural abnormalities implicating centromeric DNA. *Nat Commun*. 2018;9(1):4340. <https://doi.org/10.1038/s41467-018-06545-y>.
34. Malik HS, Henikoff S. Major evolutionary transitions in centromere complexity. *Cell*. 2009;138(6):1067–82. <https://doi.org/10.1016/j.cell.2009.08.036>.
35. Luo S, Mach J, Abramson B, Ramirez R, Schurr R, Barone P, et al. The cotton centromere contains a Ty3-gypsy-like LTR retroelement. *PLoS One*. 2012;7(4):e35261. <https://doi.org/10.1371/journal.pone.0035261>.
36. Zhang Y, Jin Y, Gong Q, Li Z, Zhao L, Han X, et al. Mechanism analysis of resistance to Verticillium dahliae in upland cotton conferred by overexpression of RPL18A-6 (Ribosomal Protein L18A-6). *Ind Crops Products*. 2019;141:111742. <https://doi.org/10.1016/j.indcrop.2019.111742>.
37. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754–60. <https://doi.org/10.1093/bioinformatics/btp324>.
38. Xu H, Luo X, Qian J, Pang X, Song J, Qian G, et al. FastUniq: a fast de novo duplicates removal tool for paired short reads. *Plos One*. 2012;7(12):e52249. <https://doi.org/10.1371/journal.pone.0052249>.
39. Wang K. Next-generation sequencing facilitates centromere position analysis of *Gossypium barbadense* and *Gossypium hirsutum*. *NCBI GEO*, GSE119184. 2019. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119184>.
40. Haixia Institute of Science and Technology. Cotton centromere study using ChIP-seq. EMBL-EBI ENA, PRJEB14368. 2017. <https://www.ebi.ac.uk/ena/browser/view/PRJEB14368>.
41. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. 2017;27(5):722–36. <https://doi.org/10.1101/gr.215087.116>.
42. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *Plos One*. 2014;9(11):e112963. <https://doi.org/10.1371/journal.pone.0112963>.
43. 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(12):1119–25. <https://doi.org/10.1038/nbt.2727>.
44. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*. 1997;268(1):78–94. <https://doi.org/10.1006/jmbi.1997.0951>.
45. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*. 2003;19(Suppl 2):ii215–25.
46. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*. 2004;20(16):2878–9. <https://doi.org/10.1093/bioinformatics/bth315>.
47. Alioto T, Blanco E, Parra G, Guigo R. Using geneid to identify genes. *Curr Protoc Bioinformatics*. 2018;64(1):e56. <https://doi.org/10.1002/cpbi.56>.
48. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*. 2008;24(24):2938–9. <https://doi.org/10.1093/bioinformatics/btn564>.
49. Keilwagen J, Hartung F, Grau J. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods Mol Biol*. 2019;1962:161–77.
50. Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Res*. 2002;12(4):656–64. <https://doi.org/10.1101/gr.229202>.
51. Hastwell AH, Gresshoff PM, Ferguson BJ. Genome-wide annotation and characterization of CLAVATA/ESR (CLE) peptide hormones of soybean (*Glycine max*) and common bean (*Phaseolus vulgaris*), and their orthologues of *Arabidopsis thaliana*. *J Exp Bot*. 2015;66(17):5271–87. <https://doi.org/10.1093/jxb/erv351>.
52. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol*. 2008;9(1):R7. <https://doi.org/10.1186/gb-2008-9-1-r7>.
53. Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res*. 2012;40(7):e49. <https://doi.org/10.1093/nar/gkr1293>.
54. Delcher AL, Salzberg SL, Phillippy AM: Using MUMmer to identify similar regions in large sequence sets. *Curr Protoc Bioinformatics* 2003, Chapter 10: Unit 10.13.
55. Sonnhammer EL, Ostlund G. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res*. 2015;43(D1):D234–9. <https://doi.org/10.1093/nar/gku1203>.
56. Chen ZJ, Sreedasyam A, Ando A, Song Q, De Santiago LM, Hulse-Kemp AM, et al. Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. *Nat Genet*. 2020;52(5):525–33. <https://doi.org/10.1038/s41588-020-0614-5>.
57. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 2003;13(9):2178–89. <https://doi.org/10.1101/gr.1224503>.
58. Nakamura T, Yamada KD, Tomii K, Katoh K. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics*. 2018;34(14):2490–2. <https://doi.org/10.1093/bioinformatics/bty121>.
59. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*. 2000;17(4):540–52. <https://doi.org/10.1093/oxfordjournals.molbev.a026334>.
60. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*. 2011;27(8):1164–5. <https://doi.org/10.1093/bioinformatics/btr088>.
61. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAXML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*. 2019;35(21):4453–5. <https://doi.org/10.1093/bioinformatics/btz305>.
62. Sanderson MJ. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*. 2003; 19(2):301–2. <https://doi.org/10.1093/bioinformatics/19.2.301>.
63. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc*. 2016;11(9):1650–67. <https://doi.org/10.1038/nprot.2016.095>.
64. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol*. 2015;16(1):259. <https://doi.org/10.1186/s13059-015-0831-x>.
65. Wang XT, Dong PF, Zhang HY, Peng C. Structural heterogeneity and functional diversity of topologically associating domains in mammalian genomes. *Nucleic Acids Res*. 2015;43(15):7237–46. <https://doi.org/10.1093/nar/gkv684>.
66. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*. 2010;38(4):576–89. <https://doi.org/10.1016/j.molcel.2010.05.004>.
67. Smyth GK. Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Edited by R. Gentleman VC, S. Dudoit, R. Irizarry, W. Huber: Springer; 2005: 397–420.
68. Zhang Y, Li T, Preissl S, Amaral ML, Grinstein JD, Farah EN, et al. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat Genet*. 2019;51(9): 1380–8. <https://doi.org/10.1038/s41588-019-0479-7>.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.