nature genetics

ARTICLES

https://doi.org/10.1038/s41588-021-00910-2

OPEN

High-quality genome assembly and resequencing of modern cotton cultivars provide resources for crop improvement

Zhiying Ma[1,3], Yan Zhang[1,3], Liqiang Wu[1,3], Guiyin Zhang[1,3], Zhengwen Sun[1,3], Zhikun Li[1,3], Yafei Jiang[2,3], Huifeng Ke[1], Bin Chen[1], Zhengwen Liu[1], Qishen Gu[1], Zhicheng Wang[1], Guoning Wang[1], Jun Yang[1], Jinhua Wu[1], Yuanyuan Yan[1], Chengsheng Meng[1], Lihua Li[1], Xiuxin Li[2], Shaojing Mo[1], Nan Wu[1], Limei Ma[1], Liting Chen[1], Man Zhang[1], Aijun Si[1], Zhanwu Yang[1], Nan Wang[1], Lizhu Wu[1], Dongmei Zhang[1], Yanru Cui[1], Jing Cui[1], Xing Lv[1], Yang Li[1], Rongkang Shi[1], Yihong Duan[1], Shilin Tian[2] and Xingfen Wang[1]

Cotton produces natural fiber for the textile industry. The genetic effects of genomic structural variations underlying agronomic traits remain unclear. Here, we generate two high-quality genomes of *Gossypium hirsutum* cv. NDM8 and *Gossypium barbadense* acc. Pima90, and identify large-scale structural variations in the two species and 1,081 G. *hirsutum* accessions. The density of structural variations is higher in the D-subgenome than in the A-subgenome, indicating that the D-subgenome undergoes stronger selection during species formation and variety development. Many structural variations in genes and/or regulatory regions potentially influencing agronomic traits were discovered. Of 446 significantly associated structural variations, those for fiber quality and Verticillium wilt resistance are located mainly in the D-subgenome and those for yield mainly in the A-subgenome. Our research provides insight into the role of structural variations in genotype-to-phenotype relationships and their potential utility in crop improvement.

As a widely cultivated fiber crop, cotton produces natural fiber for the textile industry[1]. *G. hirsutum* accounts for more than 90% of the yield in production. Thousands of improved cotton varieties have played pivotal roles in yield increases[2]. On this basis, breeders strive to create new varieties by synergistically increasing genetically complex yield and quality while obtaining resistance to numerous adversities, which is limited, however, by insufficient knowledge and understanding of the genomic basis of key agronomic traits[3]. High-quality genome assembly for modern *G. hirsutum* varieties, as well as for obsolete varieties TM-1 and ZM24 (refs. [4–6]), is crucial to breeding and biology research; however, genomic information in recently developed cottons has been limited, and genomic diversification in modern breeding process remains unclear.

G. barbadense occupies roughly 10% of the yield and affords high-quality lint fibers. To improve the fibers and disease resistance of *G. hirsutum*, a proposed approach is to transfer superior related traits from *G. barbadense* into *G. hirsutum*; however, genomic variations in *G. barbadense* compared with modern *G. hirsutum* are not clear. The identification of associated single nucleotide polymorphisms (SNPs) increases understanding of the genetic basis of cotton agricultural traits[2,7,8]. Widespread genomic structural variations, generally defined as insertion, deletion, inversion and translocation, mean that any single haplotype may be missing or contain sequence variants that are not present in most of the population[9,10]. Therefore, exploring structural variations is imperative for cotton improvement on the basis of genome assemblies and resequencing data from more accessions. Meanwhile, the genetic effects of structural variations underlying traits are less known.

In this study, we generated two high-quality reference genomes and annotations for the modern *G. hirsutum* cv. NDM8 and *G. barbadense* acc. Pima90. NDM8 is widely grown in Yellow River Valley cotton-producing areas of China, and Pima90 has served as a genetic material in molecular breeding[11–16]. Furthermore, we resequenced 1,081 worldwide *G. hirsutum* accessions, consisting of a core collection[8] plus some modern and obsolete varieties with disease resistance and glandlessness. Analyzing the two genomes and resequences showed that large-scale genomic variations occurred during breeding, providing resources for cotton crop improvement.

Results

High-quality genomes of tetraploid cottons NDM8 and Pima90. We assembled 2.29 Gb and 2.21 Gb of the NDM8 and Pima90 genomes, respectively (Table 1). To accomplish this, we obtained 205.18 Gb and 200.62 Gb long reads of NDM8 and Pima90 genomes, respectively, representing 180.38-fold coverage depth in total on the basis of single-molecule real-time (SMRT) sequencing (Supplementary Table 1). The initial assembly corrected by Illumina paired-end data (233.75-fold coverage in total) resulted in contigs with an N50 size of 15.28 Mb for NDM8 and 9.65 Mb for Pima90 (Supplementary Tables 2 and 3). Subsequently, these corrected contigs were connected to 754 superscaffolds for NDM8

[1]State Key Laboratory of North China Crop Improvement and Regulation, Key Laboratory for Crop Germplasm Resources of Hebei, Hebei Agricultural University, Baoding, China. [2]Novogene Bioinformatics Institute, Beijing, China. [3]These authors contributed equally: Zhiying Ma, Yan Zhang, Liqiang Wu, Guiyin Zhang, Zhengwen Sun, Zhikun Li, Yafei Jiang ✉e-mail: mzhy@hebau.edu.cn; zhangyan7235@126.com; tianshilin@novogene.com; cotton@hebau.edu.cn

Table 1 | Global summary of the final genome assemblies for NDM8 and Pima90

Genomic features	NDM8	Pima90
Assembled genome size (Mb)	2,291.77	2,210.14
A-subgenome (Mb)	1,438.06	1,381.46
D-subgenome (Mb)	843.88	823.21
Anchoring (%)	99.57	99.75
Number of contigs	1,030	1,160
Contig N50 (Mb)	13.15	9.24
Scaffold N50 (Mb)	107.67	102.45
Gap ratio (%)	0.003	0.06
GC content (%)	34.36	34.17
Repeat ratio (%)	62.10	61.85
Predicted PCG model number	80,124	79,613
Average gene length (bp)	2,931.27	2,894.70
Average coding sequence length per gene (bp)	1,088.90	1,094.24
Average exon number per gene	4.76	4.77
BUSCOs (%)	96.1	95.9

and 909 for Pima90 using a total of 232.90-fold 10x Genomics linked-read data (Supplementary Tables 2 and 4). Finally, we constructed chromosome-scale scaffolds using more than 125-fold Hi-C interacting unique paired-end data from each cotton genome (Extended Data Figs. 1 and 2 and Supplementary Table 2). The final assemblies included 353 scaffolds for NDM8 and 309 for Pima90, resulting in contig and scaffold N50 values of 13.15 Mb and 107.67 Mb for NDM8 and 9.24 Mb and 102.45 Mb for Pima90 (Supplementary Table 5). A total of 99.57% and 99.75% of genomes were anchored onto pseudochromosomes in NDM8 and Pima90, respectively, and the very few gaps (0.003% in NDM8 and 0.06% in Pima90) indicated the contiguity of the sequences (Supplementary Table 6). High mapping ratios (99.16% in the two genomes) and low error assembly site ratios (1.87×10^{-7} in NDM8 and 2.95×10^{-7} in Pima90) indicated the accuracy of the genomes (Supplementary Tables 7 and 8). Besides, 96.1% and 95.9% of 1,440 embryophyta Benchmarking Universal Single-Copy Orthologs (BUSCOs) present in NDM8 and Pima90, respectively, showed the integrity of the genomes (Supplementary Table 9). We compared our two genomes to a published genetic map[17], and a high consistency for each chromosome was validated for both genomes (Extended Data Figs. 3 and 4). Further, the accuracy and completeness of NDM8 assembly was confirmed by perfect alignment to 36 bacterial artificial chromosome sequences[4–6] (Supplementary Table 10). Moreover, the centromeric regions of NDM8 and Pima90 were well collinear with those of the published genomes[5] (Supplementary Tables 11 and 12). Comparing NDM8 with TM-1 (ref. [4]) and ZM24 (ref. [6]), and Pima90 with 3–79 (ref. [4]) showed a high collinearity of more than 99.69% (Supplementary Fig. 1 and Supplementary Table 13). The higher long terminal repeat (LTR) assembly index (LAI) scores[18,19] (14.2 in NDM8 and 12.1 in Pima90), as well as greater contig N50 sizes and fewer gaps in our two genomes (Supplementary Table 14) indicated that we had assembled high-quality *G. hirsutum* and *G. barbadense* genomes.

We identified 80,124 and 79,613 protein-coding gene (PCG) models in NDM8 and Pima90, respectively (Table 1 and Supplementary Tables 15 and 16), with 78,509 (98.61%) expressed PCG models in NDM8 and 78,980 (98.57%) in Pima90 on the basis of the transcriptome data from our laboratory and published data[4,5,20,21] (Supplementary Data Files 1 and 2). Compared with the PCG models from the genomes of TM-1 (refs. [4,5,20]), ZM24 (ref. [6]), Hai7124 (ref. [5]) and 3–79 (ref. [4]), and the A genome[7,22] and D genome[23], 96.98% and 97.42% of homologous PCG models had a good match, with more than 80% identity of protein sequences in NDM8 and Pima90, respectively (Supplementary Table 17). We found 1,499 and 1,267 newly predicted PCG models (identity of protein sequences <20%) in NDM8 and Pima90, respectively. Of them, 96.5% in NDM8 and 92.5% in Pima90 could be transcribed in *G. hirsutum* and *G. barbadense*, respectively (Supplementary Tables 18 and 19). Further, we discovered that NDM8 and Pima90 had lost 1,324 and 2,318 genes when compared with TM-1 (ref. [4]) and 3–79 (ref. [4]), of which 635 and 1,605 had functional annotations, respectively (Supplementary Tables 20 and 21).

We analyzed the frequency of 1,499 *G. hirsutum* newly predicted gene models in 1,081 resequenced accessions and their expression in the closely related species *G. arboreum* Shixiya1 (ref. [24]) and *G. barbadense* Pima90 and Hai7124. We found that 95.26% of the genes were harbored by at least 900 accessions (Supplementary Table 22), and 87.53% expressed in at least one variety and 100% in at least one tissue (Supplementary Table 23). Of 1,267 *G. barbadense* newly predicted genes, 90.53% were transcribed in at least one variety among Shixiya1 (ref. [24]) and five *G. hirsutum* varieties and 92.66% in at least one tissue (Supplementary Table 24).

We predicted 1,263.36 Mb and 1,204.74 Mb LTRs, which are paramount in the evolution and domestication of crops[25,26], and they covered 55.13% of NDM8 and 54.51% of Pima90 genomes (Supplementary Table 25). Of these, *Copia* was present to a much lesser extent than *Gypsy* in the NDM8 genome (17.82% versus 81.29%, $P=5.97\times10^{-27}$, Mann–Whitney *U*-test), as was also the case in Pima90 (18.14% versus 81.07%, $P=2.26\times10^{-26}$, Mann–Whitney *U*-test) (Supplementary Fig. 2 and Supplementary Table 26). We found that the number of genes with *Copia* and *Gypsy* insertions (14,900 and 14,628) was almost the same in the two genomes, and 96.69% and 95.05% of these genes were supported by transcriptome data, respectively (Supplementary Tables 26–28). The expressed gene number per *Copia* insertion was 1.84×10^{-2} and 4.68 times that per *Gypsy* insertion (3.92×10^{-3}), showing that the *Copia* impact power might be greater than that of *Gypsy*. This was further evidenced by the fact that the gene number per *Copia* insertion to exonic and promoter regions was 9.48×10^{-2} and 3.73 times that per *Gypsy* insertion (2.54×10^{-2}), which was also supported by the finding that *Copia* was markedly more active than *Gypsy* in the recent 0–1 MYA time frame[27].

We further analyzed the effects of *Copia* and *Gypsy* insertion on the gene expression of tetraploid cultivated cottons. We focused on all homologous genes between *G. barbadense* and *G. hirsutum*, and found thousands of genes diversified in *Copia* and/or *Gypsy* insertion, with 6,306 genes only in *G. barbadense* and 5,268 only in *G. hirsutum*. Additionally, *G. barbadense* had more expressed genes (5,457) but at a lower percentage (86.54%) than *G. hirsutum* (4,841, 91.89%) during fiber development. Similar trends that 82.48% genes expressed in *G. barbadense* versus 87.81% in *G. hirsutum* under *Verticillium dahliae* (Vd) stress were found. The percentage of upregulated genes (26.50% for fiber and 22.55% for Vd) was lower than that of downregulated genes (40.02% for fiber and 47.63% for Vd) in *G. barbadense*, whereas the opposite was true in *G. hirsutum* (Supplementary Tables 29–31). These findings indicated that *Copia* and *Gypsy* played important roles in agronomic character diversification during the evolution of both cotton species.

Genomic structural variations in Pima90 against NDM8. To potentially and effectively use the genomic variation of *G. barbadense* in modern *G. hirsutum* breeding programs, we aligned the Pima90 assembly onto the NDM8 genome and found high genomic diversification (Supplementary Fig. 3 and Supplementary Table 32).

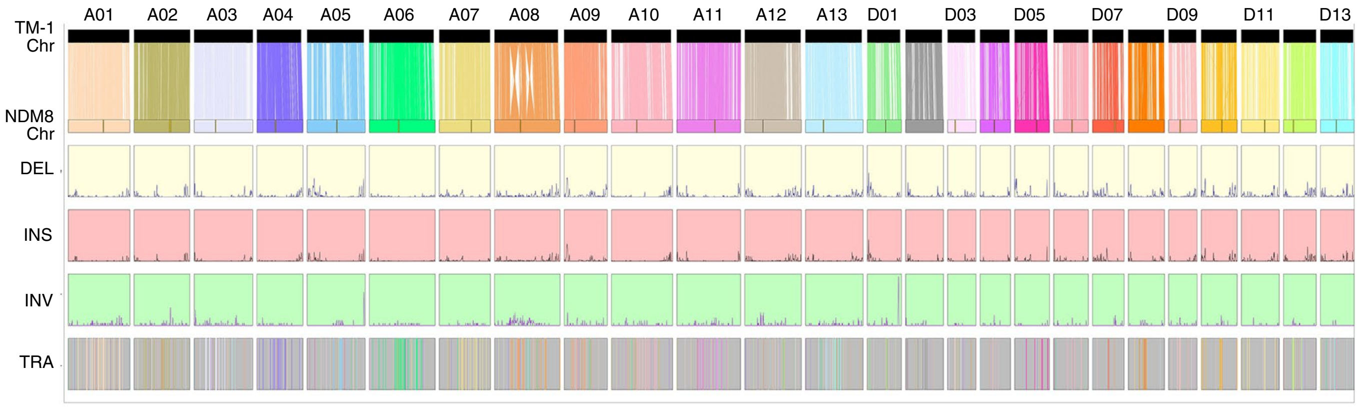


Fig. 1 | Genomic landscape of NDM8 and TM-1_HAU genomes. The vertical lines indicate the synteny between two genomes. Chr, Length of chromosome (Mb); DEL, density distribution of deletions; INS, density distribution of insertions; INV, density distribution of inversions; TRA, translocations between NDM8 and TM-1. The sliding windows are nonoverlapped with a 500-kb length.

We discovered 78,126 gene models in Pima90 homologous to 78,238 in NDM8. For the nonhomologous gene models, 1,394 were in syntenic blocks and 93 in nonsyntenic blocks (Supplementary Table 33), with 62.81% such genes expressed in several tissues (Supplementary Table 34). In total, we detected 846,363 structural variations in Pima90, with 517,230 insertions and 317,638 deletions. The top three numbers of both insertion and deletion were found on the At12, At09 and Dt11 chromosomes (t in At or Dt indicates tetraploid). Insertions and deletions ≤10 bp occupied 94.34% of the total (Supplementary Table 35). The total number of insertions and deletions in At (418,107) was almost equal to that in Dt (416,761); however, the densities of insertions (312 per megabase) and deletions (194 per megabase) in Dt were evidently higher than those in At (188 per megabase and 114 per megabase, respectively) ($P=6.43\times10^{-13}$ for insertions and $P=1.51\times10^{-13}$ for deletions, Mann–Whitney *U*-test) (Supplementary Fig. 4).

We analyzed expression changes for the insertion and deletion variant-gene pairs between *G. barbadense* and *G. hirsutum*, reflecting structural variation effect on gene expression[10]. On the basis of our transcriptome data between *G. barbadense* and *G. hirsutum*, from different fiber developmental stages, tissues (root, stem and leaf) and inoculation time-points with Vd, we found that 31,296 variant-gene pairs (the variants in genes and/or ±1 kb flanking regulatory regions) showed significantly differential expression ($\log_2$ fold-change ≥1, $P\leq0.05$) (Extended Data Fig. 5 and Supplementary Table 36), indicating that the structural variations might, to some extent, affect gene expression. Three variant-gene pairs can be exampled. Two 1-bp insertions and a 1-bp deletion located in the introns of an EXPANSIN gene *GbM_D08G1627* whose homologous protein functioned in improving fiber length (FL) and micronaire value (M)[28]. This gene was expressed in *G. barbadense* only during the fiber elongation period. Insertions of 8-bp and 1-bp were located downstream in *GbbHLH* (*GbM_A12G2140*), as were four insertions and four deletions in the introns and downstream of *GbDIR* (*GbM_A04G0106*). Both genes are positive regulators involved in lignin biosynthesis; however, excessive lignin in the cell walls of cotton fibers restricts elongation and secondary cell wall (SCW) synthesis[29,30]. The null expression of *GbbHLH* and *GbDIR* might be related to better fiber quality (Extended Data Fig. 5).

We found 5,815 variants in the exons of 5,256 genes, with 4,180 variants causing frameshift and 381 causing the gain or loss of a stop codon in Pima90 (Supplementary Table 37). A total of 3,178 variants were consistent with the transcripts from fiber, root, stem, leaf and Vd-infected tissues in *G. barbadense* and *G. hirsutum*. Among these genes, we discovered that *GbM_D13G2394*, encoding sucrose synthase (Sus), which plays a principal role in cotton fiber elongation and/or SCW synthesis[31,32], contained a transmembrane domain with a 2-bp deletion in Pima90; the *GbSus* expression was distinctly higher during fiber elongation and SCW synthesis in *G. barbadense* (Extended Data Fig. 6). This indicated that the new isoform of *GbSus* may play a crucial role in *G. barbadense* fiber length and strength. This 2-bp deletion was also identified in 3–79, Hai7124 and two *G. barbadense* introgression lines NDM373-9 and Luyuan343 (ref. [33]) with good fiber quality.

We identified 9,515 inversions with an average of 21.85 kb distributed nonrandomly across Pima90 chromosomes (Supplementary Fig. 5 and Supplementary Table 38). Of those, 6,685 and 2,830 inversions were located in At and Dt, respectively, with higher density in At (4.84×10^{-3} per kilobase) than in Dt (2.71×10^{-3} per kilobase) ($P=6.44\times10^{-9}$, Mann–Whitney *U*-test). The top three numbers of inversion were found on At06, At08 and At12, which differed from the case in 3–79 (ref. [4]). The largest inversion (585.02 kb) was located on At05, whereas the largest inversion in 3–79 (328.2 kb) was seen on Dt12. We discovered that 2,024 inversions overlapped with the exons of genes, which might lead to gene function changes (Supplementary Table 39). Additionally, we detected 1,980 translocations, of which 74.09% were interchromosomal (Supplementary Table 40).

To illustrate the potential use of *G. barbadense* germplasm in *G. hirsutum* breeding, we resequenced (30-fold) a *G. hirsutum* new line, NDM373-9, developed through backcross with the donor parent Pima90 and exhibited better Verticillium wilt (VW) resistance and fiber properties than its receptor parent *G. hirsutum* CCRI8 (Supplementary Fig. 6 and Supplementary Table 41). We found that NDM373-9 contained 171 exonic structural variations transferred from Pima90, and 34 and 12 genes with such structural variations were related to disease resistance and fiber development, respectively, as reported in previous studies (Supplementary Table 42).

Genomic structural variations in *G. hirsutum* NDM8. The high-quality genome of NDM8 allowed us to understand the genomic changes of modern *G. hirsutum* through comparison with TM-1 (ref. [4]), the two cultivars being released more than half a century apart (Supplementary Fig. 7). We identified 76,568 structural variations in NDM8 (Fig. 1 and Supplementary Table 43), including 27,708 insertions, 47,221 deletions, 808 inversions and 831 translocations. Further, we detected 28,626 consistent structural variations supported by the accessions ranging from 10 to 1,081 in the resequencing population (Supplementary Table 44).

We found that the numbers of insertions (13,985) and deletions (23,677) in At were roughly equal to those in Dt (12,705 insertions

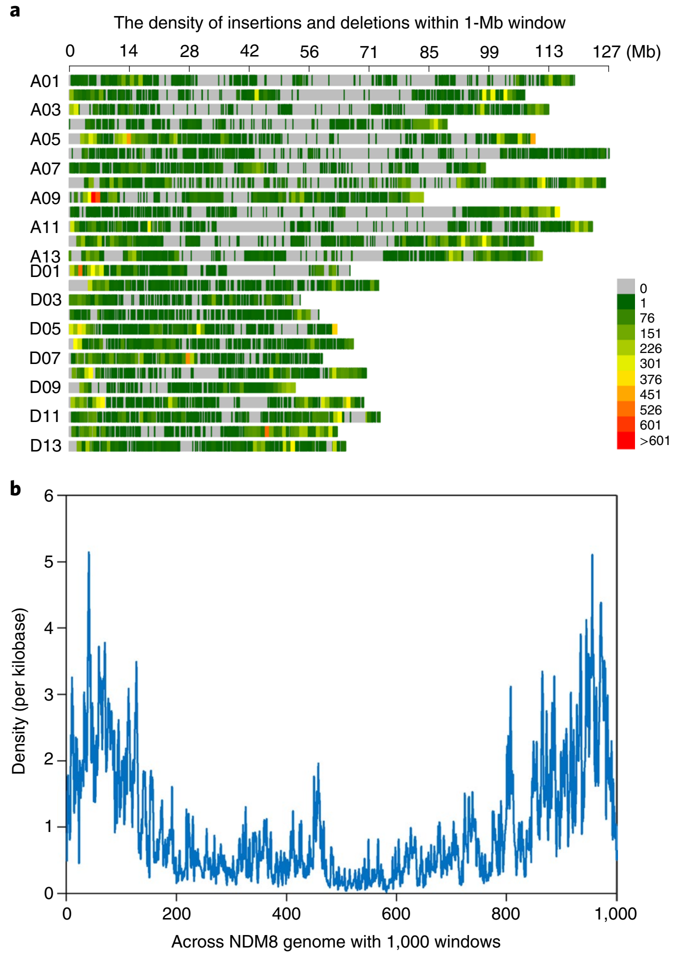


Fig. 2 | Density distribution of insertions and deletions in NDM8 genome. a, The density of insertions and deletions in a 1-Mb window of chromosomes. **b**, The density of insertions and deletions across NDM8 genome with 1,000 windows.

and 21,076 deletions); however, the densities of insertions and deletions were apparently higher in Dt ($P=1.28\times10^{-3}$ for insertions and $P=3.18\times10^{-4}$ for deletions, Mann–Whitney *U*-test) (Supplementary Fig. 8), which was also observed in the comparison of Pima90 against NDM8. We further analyzed the density of insertions and deletions across each chromosome, and observed the strongest bias within 20% of the windows near the telomeres, with a 3.71-fold ($P<10^{-6}$, permutations) increase over that in the other regions (Fig. 2). This was much higher than that of Pima90, with a 1.89-fold increase (Extended Data Fig. 7).

Furthermore, we found 603 insertions and deletions in the exons of 526 genes in NDM8 (Supplementary Table 45). Among these genes, 189 were homologous, 76 were nonhomologous and 261 were not annotated genes in the corresponding positions of TM-1, which might potentially indicate gene function changes. For example, of the 189 genes, *GhM_A02G1731* in NDM8 is homologous to the rice cinnamoyl-CoA reductase (CCR) gene that plays a role in fungal disease resistance by controlling lignin synthesis[34,35]. However, the gene in VW-susceptible TM-1 contained a 1-bp deletion in splicing site, resulting in two deletions (29 bp and 45 bp) and a truncated protein with an impaired NAD-binding domain and a lower expression level under Vd stress than that in VW-resistant NDM8 (Extended Data Fig. 8).

Of 808 inversions, the largest inversion of 1.77 Mb was located in At08, and 257 overlapped with gene models (Supplementary Tables 46 and 47). The number of inversions in At was 2.62 times that in Dt, which did not match with the fact that the genome of At was 1.70 times that of Dt, showing significantly higher density in At ($P=2.60\times10^{-5}$, Mann–Whitney *U*-test) (Supplementary Fig. 9), in contrast to the case that insertions and deletions were situated mainly in Dt in both Pima90 and NDM8. We detected that 57.52% of 831 translocations were interchromosomal (Supplementary Table 48).

Furthermore, we found 4,984 ordered genes without any structural variation (100% identity) (Supplementary Table 49) in 159,960 identical ordered synteny blocks (no gap, no mismatch and each ≥1 kb) in NDM8 (Supplementary Fig. 10), indicating that these genes might be important in maintaining fundamental biological characteristics. In addition, we made a comparison between NDM8 and ZM24 (ref. [6]), and obtained 1,393 insertions, 9,113 deletions, 243 inversions and 146 translocations (Supplementary Table 50). For the length of inversion and translocation, we found NDM8 versus ZM24 < ZM24 versus TM-1 < NDM8 versus TM-1 (Supplementary Table 51), indicating that the closer the breeding-year of two varieties were, the fewer the variations.

We analyzed the structural variations in 100 early varieties (released before 1970 and developed mainly through pedigree selection) and 100 modern varieties (released after 1990 and developed mainly through cross breeding) that were significantly improved in economic traits (Supplementary Table 52). We found that the modern varieties acquired 1,128 structural variations (in at least 51% of the varieties) compared with the early varieties during breeding (Supplementary Table 53). We found 555 and 573 acquired structural variations in At and Dt, respectively, whereas a higher density was observed in Dt (6.79×10^{-4} per kilobase) than in At (3.86×10^{-4} per kilobase) ($P=7.81\times10^{-5}$, Mann–Whitney *U*-test), implying that Dt underwent stronger selection during modern breeding.

Structural variations associated with agronomic traits in *G. hirsutum*. We explored structural variations by resequencing 1,081 *G. hirsutum* accessions (average 10.65-fold) referring to the NDM8 genome (Supplementary Table 54). On the basis of strict screening, we obtained 304,630 structural variations, including 141,145 insertions, 156,234 deletions, 39 inversions, 6,384 translocations and 828 duplications (Supplementary Table 55); 76.94% were located in intergenic regions, and the variation percentage was lower in coding sequences than in intronic regions (Supplementary Table 56). The structural variations, together with 2,970,970 SNPs and genetic kinship of all the accessions (Supplementary Fig. 11 and Supplementary Tables 57 and 58), provided broad molecular basis for cotton improvement.

So far, the genetic effects of structural variations underlying agronomically important traits remain elusive in cotton. Thus, we conducted a genome-wide association study (GWAS) for principal fiber quality and yield traits and VW resistance. The best linear unbiased prediction (BLUP) values and means for each of six traits, including FL, fiber strength (FS), M, boll weight (BW), lint percentage (LP) and seed index (SI) were calculated on the basis of phenotypic data from several environments representing years and locations (14 environments for the core collection of 419 accessions[8], eight environments for the 662 expanded accessions[36] and one environment for all 1,041 accessions in 2019). For VW resistance, the disease index (DI) of 401 accessions was determined using the high-pathogenicity Vd strain LX2-1 (ref. [37]) in a growth chamber with four independent experiments. We identified 446 structural variations significantly associated with the seven traits, of which 346 with fiber quality, 97 with yield and 3 with VW resistance (Extended Data Figs. 9 and 10 and Supplementary Data File 3). We focused on 193 structural variations simultaneously detected by both BLUP and average values (hereafter the same), and found 160 and 33 structural variations for fiber quality and yield traits, respectively. There are 29 variations in regulatory regions and 19

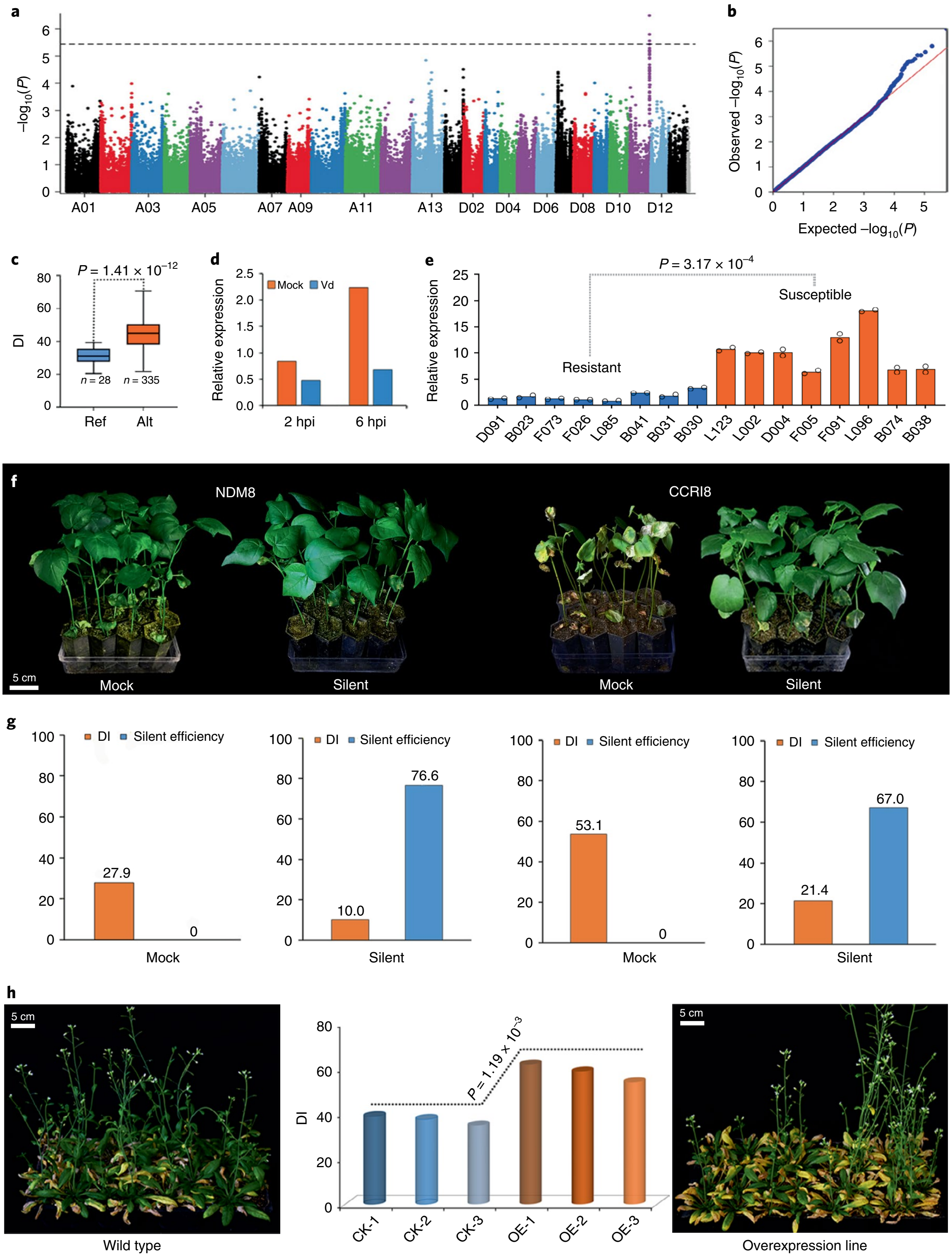


Fig. 3 | Identification of the causal gene *GhNCS* related to VW resistance on chromosome Dt11. a, Manhattan plot. Dashed line represents the significance threshold ($-\log_{10}(P) = 5.44$). We performed statistical analysis with a two-tailed Wald test. **b**, Quantile-quantile plot. **c**, Boxplot for DI on the basis of structural variation (D11: 69329075). In the box plots, the center line denotes the median, box limits are the upper and lower quartiles and whiskers mark the range of the data; *n* indicates the number of accessions with the same genotype. The difference significance was analyzed by two-tailed *t*-test. **d**, Expression level of *GhNCS* in resistant variety ND601 inoculated with Vd LX2-1. **e**, qRT–PCR analysis of *GhNCS* in eight resistant and eight susceptible varieties under Vd stress. *Ghhistone3b* was used as an internal control. The data were analyzed from the total of 16 varieties and expressed as the mean from two experiments. The difference significance was analyzed by two-tailed *t*-test. **f**, Silencing of *GhNCS* in tolerant variety NDM8 and susceptible variety CCRI8 led to obviously increased resistance compared with the mock. Scale bar, 5 cm. **g**, For each independent virus-induced gene silencing experiment, 35 cotton seedlings with higher silent efficiency were used for VW disease resistance detection. **h**, *GhNCS* overexpressed in *Arabidopsis* made transgenic plants highly susceptible compared with the wild type. Scale bar, 5 cm.

in genes that need to be the focus of functional analyses because they can directly alter the functionality of transcriptional regulatory elements and genes. The structural variations for fiber quality traits (FL, FS, M) were situated mainly in Dt (139 versus 21 in At), whereas those for yield traits (BW, LP, SI) were situated mainly in At (22 versus 11 in Dt).

For FL, which can markedly increase the economic value of end-use yarns in the textile industry, we detected the highest association peak in Dt11, where a 370-kb region (24.55–24.93 Mb) harbored 125 structural variations. Among these loci, as in NDM8, 69 and 56 increased FL significantly by 0.71–0.99 mm and by 1.00–1.19 mm, respectively (Supplementary Table 59), increasing FL from 27-mm or 28-mm grade to 29-mm grade. For the important lint yield trait LP, two structural variations in Dt03 increased LP significantly from 37.49% to 39.69% and from 37.47% to 40.00%. For VW resistance, a peak in Dt11 (69.00–69.33 Mb) with three structural variations caused a DI decline of more than 13.6 in the genotype, the same as the resistant NDM8, shifting the disease reaction from susceptible (DI = 44.5–45.2) to tolerant (DI = 30.9–31.1) (Fig. 3a–c).

We identified 907 candidate genes for fiber quality and yield traits and 60 for VW resistance on the basis of a linkage disequilibrium decay value of 325 kb (Supplementary Fig. 12). We found 84.23% genes expressed at the fiber developmental stages of *G. hirsutum*, of which 305 had structural variations in genes and regulatory regions (Supplementary Data File 3), implying that these genes might potentially influence fiber quality and yield. Moreover, we found that four deletions in the 5′ untranslated region (UTR), intronic and 3′ UTR of *GhM_D11G2206* were significantly associated with FL. This gene was the same as the validated *GhFL2* in our previous study[8].

To validate the reliability of GWAS results for significant hits, we chose the gene *GhM_D11G3743* associated with two structural variations in Dt11. This gene encodes (*S*)-norcoclaurine synthase, designated as *GhNCS*, and is a member of the pathogenesis-related 10/Bet v1 protein family[38] whose function in cotton disease resistance is unclear. qRT–PCR assays showed that *GhNCS* expression was downregulated under Vd stress compared with mock and significantly lower amounts in eight resistant varieties (reference genotype) than in eight susceptible varieties (alternative genotype) (Fig. 3d,e and Supplementary Table 60). Silencing *GhNCS* in cotton resulted in resistance enhancement in both susceptible and resistant varieties, making the highly susceptible variety CCRI8 (DI = 53.1) tolerant (DI = 21.4) and the tolerant variety NDM8 (DI = 27.9) resistant (DI = 10.0) (Fig. 3f,g). Nevertheless, overexpression of *GhNCS* in *Arabidopsis* made the transgenic plants highly susceptible (DI = 58.1) compared with the wild type (DI = 38.1) (Fig. 3h). These results indicate that *GhNCS* is a plausible causal gene controlling VW resistance and that the associated structural variations are reliable.

Discussion

In the present work, we completed two new high-quality assemblies of modern *G. hirsutum* cv. NDM8 and *G. barbadense* acc. Pima90, and detected many interspecific and intraspecific genomic variations. More and larger inversions occurred in the A-subgenome of *G. hirsutum*, which was similar to the recent reports[6,20,39]; however, the D-subgenome acquired more insertions and deletions than the A-subgenome during modern breeding. The density of insertions and deletions across each chromosome showed the strongest bias near the telomeres, similar to what has been reported in the human genome[10]. These will enhance the genomic resources for cotton improvement and provide insight into species formation and variety development.

There are several reports about the genomic diversity of *Gossypium* allopolyploid species on the basis of sequencing *G. hirsutum* TM-1, ZM24, *G. barbadense* Hai7124, 3–79, *G. tomentosum*, *G. mustelinum* and *G. darwinii*[4–6,39,40] and resequencing large-scale accessions[8,41]. On the basis of the sum of the gene number in each gene family counting by the priority in 3–79 > TM-1_HAU > Hai7124 > TM-1_ZJU > ZM24 > TM-1_CRI tetraploid cottons, we found that 15,973 genes might actually belong to duplicates and/or alleles of some genes, and 80,992 were nonredundant in the six genomes (Supplementary Table 61), which provides new information for plant genome researchers.

We found that a 2-bp deletion in *GbSus* in the D-subgenome of Pima90 (also existed in 3–79 and Hai7124) diverged from species formation because the deleted AC bases could be detected in the D-subgenome of NDM8, TM-1 and ZM24 and traced in the ancestral diploid species *G. ramondii* (Extended Data Fig. 6). Similarly, a 1-bp insertion in *CCR* in the A-subgenome of NDM8 could be found in Pima90, 3–79, Hai7124 and ZM24 and traced in the ancestral diploid species *G. arboreum* Shixiya1 (Extended Data Fig. 8). We inferred that NDM8 regained the insertion from its pedigree ancestral varieties, excluding TM-1 and its selections, during artificial recombination in breeding.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00910-2.

Received: 26 November 2019; Accepted: 8 July 2021;
Published online: 9 August 2021

References

1. Chen, Z. J. et al. Toward sequencing cotton (*Gossypium*) genomes. *Plant Physiol.* **145**, 1303–1310 (2007).
2. Fang, L. et al. Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. *Nat. Genet.* **49**, 1089–1098 (2017).
3. International Wheat Genome Sequencing Consortium. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, eaar 7191 (2018).
4. Wang, M. J. et al. Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. *Nat. Genet.* **51**, 224–229 (2019).
5. Hu, Y. et al. *Gossypium barbadense* and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton. *Nat. Genet.* **51**, 739–748 (2019).
6. Yang, Z. E. et al. Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. *Nat. Commun.* **10**, 2989 (2019).
7. Du, X. M. et al. Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. *Nat. Genet.* **50**, 796–802 (2018).
8. Ma, Z. Y. et al. Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. *Nat. Genet.* **50**, 803–813 (2018).
9. Mills, R. E. et al. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
10. Audano, P. A. et al. Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663–675 (2019).
11. He, D. H. et al. QTL mapping for economic traits based on a dense genetic map of cotton with PCR-based markers using the interspecific cross of *Gossypium hirsutum* × *Gossypium barbadense*. *Euphytica* **153**, 181–197 (2007).
12. Liu, X. et al. Identification and expression profile of *GbAGL2*, a C-class gene from *Gossypium barbadense*. *J. Biosci.* **34**, 941–951 (2009).
13. Zhang, Y. et al. Targeted transfer of trait for Verticillium wilt resistance from *Gossypium barbadense* into *G. hirsutum* using SSR markers. *Plant Breed.* **135**, 476–482 (2016).
14. Yang, X. L. et al. Mapping QTL for cotton fiber quality traits using simple sequence repeat markers, conserved intron-scanning primers, and transcript-derived fragments. *Euphytica* **201**, 215–230 (2015).

15. Zhang, Y. et al. Histochemical analyses reveal that stronger intrinsic defenses in *Gossypium barbadense* than in *G. hirsutum* are associated with resistance to *Verticillium dahliae*. *Mol. Plant Microbe Interact.* **30**, 984–996 (2017).
16. Tang, M. et al. Rapid evolutionary divergence of *Gossypium barbadense* and *G. hirsutum* mitochondrial genomes. *BMC Genomics* **16**, 770 (2015).
17. Wang, S. et al. Sequence-based ultra-dense genetic and physical maps reveal structural variations of allopolyploid cotton genomes. *Genome Biol.* **16**, 108 (2015).
18. Qu, S. J. et al. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
19. Grover, C. E. et al. The Gossypium longicalyx genome as a resource for cotton breeding and evolution. *G3 (Bethesda)* **10**, 1457–1467 (2020).
20. Zhang, T. Z. et al. Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat. Biotechnol.* **33**, 531–537 (2015).
21. Liu, X. et al. *Gossypium barbadense* genome sequence provides insight into the evolution of extra-long staple fiber and specialized metabolites. *Sci. Rep.* **5**, 14139 (2015).
22. Li, F. G. et al. Genome sequence of the cultivated cotton *Gossypium arboreum*. *Nat. Genet.* **46**, 567–572 (2014).
23. Wang, K. B. et al. The draft genome of a diploid cotton *Gossypium raimondii*. *Nat. Genet.* **44**, 1098–1104 (2012).
24. Wang, K. et al. Multi-strategic RNA-seq analysis reveals a high-resolution transcriptional landscape in cotton. *Nat. Commun.* **10**, 4714 (2019).
25. Feschotte, C., Jiang, N. & Wessler, S. R. Plant transposable elements: where genetics meets genomics. *Nat. Rev. Genet.* **3**, 329–341 (2002).
26. Xiao, H., Jiang, N., Schaffner, E., Stockinger, E. J. & van der Knaap, E. A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science* **319**, 1527–1530 (2008).
27. Li, F. et al. Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nat. Biotechnol.* **33**, 524–530 (2015).
28. Bajwa, K. S. et al. Stable transformation and expression of *GhEXPA8* fiber expansin gene to improve fiber length and micronaire value in cotton. *Front. Plant Sci.* **6**, 838 (2015).
29. Gao, Z. Y. et al. *GhbHLH18* negatively regulates fiber strength and length by enhancing lignin biosynthesis in cotton fibers. *Plant Sci.* **286**, 7–16 (2019).
30. Davin, L. B. & Lewis, N. G. Lignin primary structures and dirigent sites. *Curr. Opin. Biotechnol.* **16**, 407–415 (2005).
31. Ruan, Y. L., Llewellyn, D. J. & Furbank, R. T. Suppression of sucrose synthase gene expression represses cotton fiber cell initiation, elongation, and seed development. *Plant Cell* **15**, 952–964 (2003).
32. Brill, E. et al. A novel isoform of sucrose synthase is targeted to the cell wall during secondary cell wall synthesis in cotton fiber. *Plant Physiol.* **157**, 40–54 (2011).
33. Wang, F. R. et al. Identification of candidate genes for key fibre-related QTLs and derivation of favourable alleles in *Gossypium hirsutum* recombinant inbred lines with *G. barbadense* introgressions. *Plant Biotechnol. J.* **18**, 707–720 (2020).
34. Kawasaki, T. et al. Cinnamoyl-CoA reductase, a key enzyme in lignin biosynthesis, is an effector of small GTPase Rac in defense signaling in rice. *Proc. Natl Acad. Sci. USA* **103**, 230–235 (2006).
35. Bart, R. S., Chern, M., Vega-Sánchez, M. E., Canlas, P. & Ronal, P. C. Rice *Snl6*, a cinnamoyl-CoA reductase-like gene family member, is required for NH1-mediated immunity to *Xanthomonas oryzae* pv. oryzae. *PLoS Genet.* **6**, e1001123 (2010).
36. Sun, Z. W. et al. Genome-wide association study discovered genetic variation and candidate genes of fibre quality traits in *Gossypium hirsutum* L. *Plant Biotechnol. J.* **15**, 982–996 (2017).
37. Zhang, Y. et al. The cotton laccase gene *GhLAC15* enhances Verticillium wilt resistance via an increase in defence-induced lignification and lignin components in the cell walls of plants. *Mol. Plant Pathol.* **20**, 309–322 (2018).
38. Lee, E. J. & Facchini, P. Norcoclaurine synthase is a member of the pathogenesis-related 10/Bet v1 protein family. *Plant Cell* **22**, 3489–3503 (2010).
39. Chen, Z. J. et al. Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. *Nat. Genet.* **52**, 525–533 (2020).
40. Huang, G. et al. Genome sequence of *Gossypium herbaceum* and genome updates of *Gossypium arboreum* and *Gossypium hirsutum* provide insights into cotton A-genome evolution. *Nat. Genet.* **52**, 516–524 (2020).
41. He, S. P. et al. The genomic basis of geographic differentiation and fiber improvement in cultivated cotton. *Nat. Genet.* **53**, 916–924 (2021).

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

Methods

Plant material and resequencing. *G. hirsutum* cv. NDM8 and *G. barbadense* acc. Pima90 (self-pollinated for more than ten generations) were selected for genome sequencing because of their important roles in cotton genetic research and breeding. NDM8 was released in 2006, with high yield, good fiber properties and resistance to Fusarium wilt and VW. Pima90 is selected from Pima cotton. A total of 1,081 *G. hirsutum* accessions from China and other countries were used for resequencing according to our previous description[8] (Supplementary Table 55). After germination, five full seeds of each accession were planted in pots with vermiculite and cultured at 27 °C in a growth chamber. After two cotyledons spread, the cotyledons of a single seedling were harvested and frozen immediately in liquid nitrogen for the extraction of genomic DNA.

Genomic DNA for PacBio. Total genomic DNA from two cottons, NDM8 and Pima90, was extracted for sequencing using the CTAB method. To construct sequencing libraries, genomic DNA was fragmented by g-TUBE, centrifuged at 2,000 r.p.m. for 2 min, and treated with end-repair, adapter ligation and exonuclease digestion as recommended by Pacific Biosciences. DNA fragments at 10–50 kb were selected by Blue Pippin electrophoresis (Sage Sciences). DNA libraries were sequenced on the PacBio Sequel platform (Pacific Biosciences) with Sequel Sequencing chemistry v.3.0. A total of 21 SMRT cells were sequenced for NDM8 producing 205.41 Gb of polymerase reads and 27 cells for Pima90 producing 200.82 Gb of raw data. For the PacBio data, subreads were filtered with the default parameters, and the N50 length of long subreads reached 19.84 kb and 18.82 kb in NDM8 and Pima90, respectively.

Illumina paired-end sequencing. Genomic DNA of each accession was extracted (1.5 µg per sample) and used as input material for DNA sample preparation. Sequencing libraries were generated using a TruSeq Nano DNA HT Sample Preparation Kit (Illumina) following the manufacturer's instructions, and index codes were added to attribute sequences to each sample. Briefly, the DNA samples were fragmented by sonication to short inserts (350 bp), and the DNA fragments were then end-polished, A-tailed and ligated with the full-length adapters for Illumina sequencing with further PCR amplification. Finally, PCR products were purified (AMPure XP), and the libraries were analyzed for size distribution using an Agilent 2100 Bioanalyzer and quantified using real-time PCR.

10x Genomics library construction, sequencing and extension scaffold. The GemCode Instrument from 10x Genomics was used for DNA sample preparation, indexing and barcoding. Around 1 ng of input DNA with a 50-kb length was used for the GEM reactions during PCR, and 16-bp barcodes were introduced into droplets. Then, the droplets were fractured following purification of the intermediate DNA library. Next, we sheared DNA into 500-bp fragments for constructing libraries, which were finally sequenced on NovaSeq.

Hi-C library construction and sequencing. We constructed Hi-C libraries from cotton leaves of NDM8 and Pima90. The leaves were fixed with formaldehyde and lysed. After that, we digested the cross-linked DNA with *Hin*dIII. Sticky ends were biotinylated and proximity-ligated to form chimeric junctions. They were then enriched and physically sheared into fragments of 300–500 bp. The chimeric fragments representing the original cross-linked long-distance physical interactions were processed into paired-end sequencing libraries. Finally, 150-bp paired-end sequences were produced on the Illumina platform[42].

Sequence quality checking and filtering. We used strict filters to avoid reads with artificial bias for Illumina paired-end sequences, 10x Genomics linked reads and Hi-C data. First, low-quality paired reads (reads with ≥10% unidentified nucleotides (*N*); >10 nt aligned to the adapter, allowing ≤10% mismatches; >50% bases having phred quality <5 and putative PCR duplicates generated in the library construction process), which resulted mainly from base-calling duplicates and adapter contamination, were removed. Consequently, we obtained 32.24 Tb of high-quality data for collection, extension, chromosome-scale scaffolds and large-scale population analysis.

Hi-C reads mapping, filtering and generation of contact matrices. Initial Hi-C data analyses including read mapping, filtering and bias correction were conducted by Hiclib (https://github.com/mirnylab/hiclib-legacy). High-quality paired-end reads were mapped to the two genomes by Bowtie2 (ref. [43]) (with the 'very-sensitive' option) through iterative mapping. Mapped reads were filtered using Hiclib[44] with default parameters, discarding the invalid self-ligated and unligated fragments and PCR artifacts. Valid Hi-C read pairs harbored more intrachromosomal (cis) interactions than interchromosomal (trans) interactions. Normalized interaction matrices were generated at four resolutions from low to high: 1 Mb, 500 kb, 100 kb and 40 kb.

Genome assembly. First, the package 'daligner' of the FALCON assembler[45] was used to self-correct PacBio long reads using the PacBio short reads less than 5,000 bp. Then, contigs of the two cottons were assembled using the package FALCON assembler on the basis of the error-corrected reads. The overlapped read pairs were used to construct a directed string graph following Myers' algorithm. Contigs were constructed by finding the paths from the string graph. The preceding assemblies were polished by the consensus–calling algorithm Quiver[46]. We mapped Illumina paired-end reads to the contig assemblies and corrected them using the Pilon pipeline[47]. The corrected contigs were further connected to generate superscaffolds by 10x Genomics linked-read data using fragScaff software[48]. Linkage information of superscaffolds was obtained by aligning high-quality Hi-C data to the preceding assemblies using Bowtie2 software. Chromosome-scale scaffolds were anchored by linkage information, restriction enzyme site, and string graph formulation with the package LACHESIS[49]. Hi-C data were mapped to chromosome-scale scaffolds to assess the quality of assemblies using HiC-Pro software[50] (v.2.10.0). The placement and orientation errors exhibiting obvious discrete chromatin interaction patterns were adjusted manually.

Assessment of genome assembly quality. To validate the single-base accuracy of the genome assemblies, we realigned the high-quality 350-bp paired-end reads to the assemblies with BWA software[51]. More than 99.67% of the genome having a coverage depth ≥10 indicated an extremely high sequencing depth over the whole genome. We conducted variant calling with SAMtools[52] and obtained homozygous SNP (that is, error assembly site). We used BUSCO analysis[53] to assess genome completeness by searching against the embryophyta BUSCO (v.3.0).

Genome repeat annotation. The repetitive sequences in the cotton genome were identified by a combination of homology searching and ab initio prediction. For homology-based prediction, we used RepeatMasker[54] and RepeatProteinMask to search against Repbase. For ab initio prediction, we used Tandem Repeats Finder[55], LTR FINDER[56], PILER[57] and RepeatScout[58] with default parameters. The code used for the genome annotations of repetitive elements is deposited in the Zenodo DOI-minting repository[59].

Structural annotation of genes. Gene prediction was conducted through a combination of homology- and ab initio–based methods and by incorporating evidence from transcriptions. Proteins of plants, including *Gossypium hirsutum* (http://cotton.hzau.edu.cn/EN/download.php, http://ibi.zju.edu.cn/cotton/), *Gossypium barbadense* (http://cotton.hzau.edu.cn/EN/download.php, http://ibi.zju.edu.cn/cotton/), *Gossypium raimondii* (https://phytozome.jgi.doe.gov/pz/portal.html#!bulk?org=Org_Graimondii), *Gossypium arboreum* (ftp://bioinfo.ayit.edu.cn/downloads/), *Theobroma cacao* (GCF_000208745.1), *Oryza sativa* (R498, IGDBV2), *Glycine max* (GWHAAEV00000000), *Populus trichocarpa* (GCF_000002775.4) and *Arabidopsis thaliana* (GCA_000001735.1) were used as queries to search against two cotton genomes using TBLASTN[60] with an E-value cutoff of 1×10^{-7}. The BLAST hits were conjoined by Solar software[61]. Then, we removed conjoined query hits with <25% coverage and merged two hits with >50% overlap in length. Subsequently, GeneWise[62] was used to predict the exact gene structure of the corresponding genomic region on each conjoined hit. Homology predictions were denoted as 'Homology-set'.

For transcription evidence, RNA-seq data of the four cotton tissues root, stem, leaf, fiber and public data from nine tissues[20] were used. Illumina RNA-seq data were assembled by Trinity[63], and full-length nonchimeric transcripts were obtained using IsoSeq3 pipeline (https://anaconda.org/bioconda/isoseq3) on the basis of PacBio sequences. Subsequently, these transcripts were aligned against two cotton genomes by the Program to Assemble Spliced Alignment (PASA)[64] with default parameters. Valid transcript alignments were clustered on the basis of genome mapping location and assembled into gene structures. Gene models created by PASA were denoted as PASA Trinity set (PASA-T-set). In addition, Illumina RNA-seq reads were mapped to the genome using Tophat[65] to identify putative exonic regions and splicing junctions, and then Cufflinks[66] was used to assemble the mapped reads into gene models (Cufflinks-set).

We performed ab initio prediction for coding regions in the repeat-masked genome using Augustus[67], GeneID[68], GenScan[69], GlimmerHMM[70] and SNAP[71]. Specifically, GeneID and GenScan with the self-trained model parameters (*A. thaliana*) were used to predict two masked cotton genomes; Augustus, SNAP and GlimmerHMM were trained by PASA-H-set gene models; Augustus, SNAP and GlimmerHMM were used to predict two masked cotton genomes.

Gene models generated from all the methods were integrated by EvidenceModeler[72]. Weights for each type of evidence were set as follows: PASA-T-set > Homology-set > Cufflinks-set > Augustus > GeneID = SNAP = GlimmerHMM = GenScan. A weighted and nonredundant gene set were further revised by PASA2 to generate untranslated regions and alternative splicing variation information. The code used for the genome annotations of gene structures is deposited in the Zenodo DOI-minting repository[59].

Functional annotation of protein-coding genes. Gene functions of PCGs were annotated by searching for functional motifs and domains of genes and the possible biological processes in the databases SwissProt[73], Pfam[74], NR database (from National Center for Biotechnology Information (NCBI)), Gene Ontology[75] and Kyoto Encyclopedia of Genes and Genomes[76].

Estimating the theoretical gene number of tetraploid cotton genome. We carried out gene orthologous cluster analysis of tetraploid cottons using the published cotton PCG models from the genomes of 3–79_HAU, TM-1_HAU, Hai7124_ZJU, TM-1_ZJU, ZM24_CRI and TM-1_CRI. Specifically, for genes with alternative splicing sites, we chose the longest translation to represent each gene and filtered genes with fewer than 50 amino acids. To build a graph of PCGs, all-against-all BLASTP was used to determine similarities between all genes in the six cottons with an E-value of 1×10^{-7}. Subsequently, we conjoined fragmental alignments to cluster gene pairs by the OrthoMCL[77] method with the parameter '-inflation 1.5'. Finally, we obtained 47,147 gene clusters. The largest theoretical gene resource is the sum of the largest number of genes in each gene family counting by the priority (3–79_HAU > TM-1_HAU > Hai7124_ZJU > TM-1_ZJU > ZM24_CRI > TM-1_CRI). Next, to filter duplicates and/or alleles of some genes between/within six tetraploid cottons, we extracted alignment pairs from any pair of genomes and restricted a maximum of five hits per protein sequence to serve as input for the MCScanX algorithm[78] that was used to detect high-confidence collinear blocks of coding genes and identify orthologous gene pairs. Finally, we filtered 15,973 genes that might actually belong to duplicates and/or alleles of some genes.

Synteny gene identification. We identified synteny blocks through genome alignment applying the MUMmer program[79] (v.3.2) with the command 'nucmer --mum --maxgap=500 --mincluster=1000'. Meanwhile, protein sequences were compared for identifying homologous genes by using all-by-all BLASTP[60] (v.2.2.26; by E-value $\leq 1\times10^{-7}$ and identity ≥20%). Subsequently, we identified the homologous genes in one-to-one genomic synteny blocks through intersection using BEDTools[80] (v.2.27). Finally, we defined those homologous gene patterns to be ordered genes.

Genomic variation detection. To compare two genomes, we used smartie-sv software[81] to detect insertions and deletions. To filter out spurious insertions and deletions, we separately aligned the reads onto two genomes using BWA[51], and calculated the read coverage for each candidate variant. Then, different criteria were used to validate the candidates ≤50 bp and those >50 bp. Some candidates (≤50 bp) were supported by more than three gapped aligned reads and their predicted breakpoints and/or genotypes were perfectly consistent with the aligned reads. The other candidates (>50 bp) should have significant differences in S/P ratio (that is, the number of aligned single-end reads versus the number of aligned paired-end reads) between two genomes ($P < 0.05$, Fisher's exact test) and were more than three times the s.d. of the insert size in length. We detected inversion and translocation on the basis of the reverse-pattern and nonsequential-pattern synteny of the two genomes, respectively.

For population genomic variations, we separately aligned the individual sequence onto the NDM8 genome using BWA and Sentieon softwares[82] to detect SNPs (MAF ≥ 0.05, missing ratio ≤0.2, depth ≥3) and small structural variations including insertions and deletions ≤250 bp (MAF ≥ 0.05, missing ratio ≤0.2, depth ≥3), respectively. Subsequently, we identified potential large structural variations using an SVMerge pipeline[83] by integrating calls from the packages LUMPY[84] and Breakdancer[85]. Specifically, we first applied the packages LUMPY and Breakdancer to identify insertions, deletions, duplications, inversions and translocations for 1,081 accessions. The raw merged dataset contained insertions, deletions, inversions and duplications but not translocations. Next, each structural variation call was evaluated by local assembly using Velvet[86], and then contig alignments were computationally parsed to determine if there was supporting evidence for the structural variation, and to localize the breakpoints of the structural variation. On the basis of the above pipeline, the above four kinds of structural variation call sets were obtained. For translocation, we considered the calls supported by both LUMPY and Breakdancer to be reliable. Finally, for the whole set, we merged the calls of all individuals to a nonredundant set and ensured that each call had at least ten accessions to support. We constructed the phylogenetic tree applying TreeBest software (v.1.9.2).

To identify NDM373-9 fragments transferred from *G. barbadense* Pima90, we separately mapped the resequences of NDM373-9 and CCRI8 to the NDM8 reference genome and detected the specific structural variations of NDM373-9. Finally, we obtained the overlapped structural variations by comparing these structural variations to the specific structural variations of Pima90 against the NDM8 genome.

GWAS analysis. As we know, At08 possessed abundant inversions[6] that might interfere with the accuracy of GWAS. Thus, we used 277,292 structural variations excluding those located on At08 and phenotypic data to perform GWAS for the seven traits, including FL, FS, M, BW, LP, SI and VW resistance.

For fiber quality and yield traits, we used the data in our previous research, 12 environments for the core collection of 419 accessions[8] and eight environments for the 662 expanded accessions[36,87]. In addition, we newly obtained fiber quality trait data of 419 accessions collected from the Hainan breeding nursery in 2016 and 2017 and fiber quality and yield trait data for all the above 1,041 accessions from the Qingxian breeding nursery in 2019. The means and BLUP[88] were used to perform GWAS. The BLUP was calculated with lme4 packages (1.1–23) in R (v.3.6.3), and the formula was as follows:

$$Y = \mu + \text{Line} + \text{Loc} + (\text{Line} \times \text{Loc}) + (\text{Rep} \times \text{Loc}) + \epsilon$$

where Y, μ, Line and Loc represent phenotype, intercept, variety effects and environmental effects, respectively. Rep means different repetitions and ε represents random effects. Line × Loc represents the interaction between variety and environment, and Rep × Loc represents the interaction between repetition and environment.

For VW resistance evaluation, we used the high-pathogenicity strain LX2-1 to inoculate 401 out of 1,081 accessions. For each accession, we performed four independent experiments in growth chamber; 35 seedlings were analyzed in each experiment for each accession. The susceptible variety Jimian11 and the resistant variety ND601 were used as controls to monitor the accuracy of disease determination. Symptom development was recorded at 20 days post inoculation (dpi) and categorized into five grades recorded as 0 to 4. The DI was calculated according to a previous method[37].

Association analysis was conducted with the genome-wide efficient mixed-model association (GEMMA) software package[89]. The top three principal components (PCs) were used to build up the S matrix for population–structure correction. The matrix of simple matching coefficients was used to build up the K matrix. The genome-wide significance threshold was set as $P = 1/n$ (n, total number of structural variations).

RNA extraction and qRT–PCR analysis. Total RNA was extracted using the EASYspin Plus Plant RNA Kit (Aidlab Biotech) according to the manufacturer's protocols. cDNA was generated with a PrimeScript RT Reagent Kit with gDNA Eraser (TaKaRa). We performed qRT–PCR with a SYBR Premix DimerEraser (Perfect Real Time) (TaKaRa). *Ghhistone3b* was used to normalize all qRT–PCR data. The relative expression was calculated using the $2^{-\Delta\Delta Ct}$ method[90]. The primers used for gene expression analysis were listed in Supplementary Table 60.

Generation of transgenic *Arabidopsis* and disease assays. For *GhM_D11G3743* overexpression, full-length open reading frame was amplified by PCR using cDNA synthesized from RNA that was isolated from seedlings of NDM8. The amplified product was further cloned into the pGreen vector under the control of the cauliflower mosaic virus 35S promoter. The transformed seedlings were identified on the basis of Basta screening and PCR detection. T_3 seeds of transgenic lines were used for phenotypic analyses. *Arabidopsis* plants (20 d old) were inoculated with Vd as previously described[37]. Disease development was monitored for up to 28 dpi and DI was calculated according to a previous description[37].

Virus-induced gene silencing in cotton and pathogen inoculation. The gene-specific region for *GhM_D11G3743* (*GhNCS*) was amplified as a template and cloned into the pTRV2 vector. The resulting pTRV2 construct was coinfiltrated with pTRV1 via *Agrobacterium tumefaciens* GV3101 into cotton seedlings of resistant NDM8 and susceptible CCRI8, through syringe inoculation when the cotyledons opened[91]. Plants coinfiltrated with empty pTRV2 and pTRV1 were used as mock controls. After 2 weeks, the plants were inoculated with a Vd spore suspension (around 1×10^7 conidia per milliliter). We performed the experiments with at least 35 seedlings per treatment and repeated them twice. We determined the silent efficiency of cotton by using mix sample with all the treated seedlings. The DI was calculated as above. Primers used for construction of a VIGS vector are listed in Supplementary Table 60.

Statistical analysis. We performed permutation tests 1,000,000 times on the basis of the count density of the structural variations through dividing each chromosome into 1,000 sliding windows. The Mann–Whitney U-test was used to perform a statistical analysis on the densities of structural variations and LTRs. SPSS22 was used for statistical analysis of the phenotypic traits. We performed one-way analysis of variance, and the significance level was set at $P = 0.05$ or 0.01. In transcriptome analyses, the RPKM values of genes from each sample were calculated with Cufflinks (v.2.1.1)[66]. Two-tailed Student's t-tests were used to compare *GhNCS* expression levels between resistant and susceptible varieties, the DI values of the silent and mock plants and the DI values of overexpression *Arabidopsis* and mock plants.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The raw sequencing data and transcriptome data of NDM8 and Pima90, and the resequencing data of 1,081 accessions are deposited in the NCBI Sequence Read Archive under the BioProject accession number PRJNA680449. The two cotton assemblies have been deposited in NCBI GenBank under the accession numbers JAHMMW000000000 and JAHMMX000000000. The versions described in this paper are version JAHMMW000000000.1 and JAHMMX000000000.1. The relevant data are also deposited in the CottonGen database https://www.cottongen.org/ (the assemblies and gene annotations) and are available at the website http://cotton.hebau.edu.cn/Data%20Download.html (the assemblies, gene annotations, structural variations and phenotypic data).

Code availability

Code used for the genome annotations of repetitive elements and gene structures are deposited in Zenodo DOI-minting repository with the https://doi.org/10.5281/zenodo.4851529.

References

42. Belton, J. M. et al. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).
43. Langmead, B. & Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
44. Imakaev, M. et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012).
45. Pendleton, M. et al. Assembly and diploid architecture of an individual human genome via single molecule technologies. *Nat. Methods* **12**, 780–786 (2015).
46. Chin, C. S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
47. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
48. Adey, A. et al. In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Res.* **24**, 2041–2049 (2014).
49. Bruton, J. N. et al. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
50. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
51. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
52. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
53. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
54. Bergman, C. M. & Quesneville, H. Discovering and detecting transposable elements in genome sequences. *Brief. Bioinform.* **8**, 382–392 (2007).
55. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
56. Xu, Z. & Wang, H. LTR FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
57. Edgar, R. C. & Myers, E. W. PILER: identification and classification of genomic repeats. *Bioinformatics* **213**, i152–i158 (2003).
58. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
59. FionaJ1. FionaJ1/NG-A53330-code: NG-A53330-code. *Zenodo* https://doi.org/10.5281/zenodo.4851529 (2021).
60. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
61. Yu, X. J., Zheng, H. K., Wang, J., Wang, W. & Su, B. Detecting lineage-specific adaptive evolution of brain-expressed genes in human using rhesus macaque as outgroup. *Genomics* **88**, 745–751 (2006).
62. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
63. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
64. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
65. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
66. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
67. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, ii215–ii225 (2003).
68. Guigó, R. Assembling genes from predicted exons in linear time with dynamic programming. *J. Comput. Biol.* **5**, 681–702 (1998).
69. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
70. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
71. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
72. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
73. Apweiler, R. et al. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2004).
74. Finn, R. D. et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285 (2016).
75. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45**, D331–D338 (2017).
76. Kanehisa, M. et al. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42**, D199–D205 (2014).
77. Li, L., Stoeckert, C. J. Jr. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
78. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
79. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
80. Quinlan, A. R. BEDTools: the Swiss-Army tool for genome feature analysis. *Curr. Protoc. Bioinforma.* **47**, 11.12.1–11.12.34 (2014).
81. Kronenberg, Z. N. et al. High-resolution comparative analysis of great ape genomes. *Science* **360**, eaar 6343 (2018).
82. Freed, D., Aldana, R., Weber, J. A. & Edwards, J. S. The Sentieon Genomics Tools – a fast and accurate solution to variant calling from next-generation sequence data. Preprint at *bioRxiv* https://doi.org/10.1101/115717 (2017).
83. Wong, K., Keane, T. M., Stalker, J. & Adams, D. J. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol.* **11**, R128 (2010).
84. Ryan, M. L., Colby, C., Aaron, R. Q. & Ira, M. H. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
85. Chen, K. et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6**, 677–681 (2009).
86. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
87. Sun, Z. W. et al. A genome-wide association study uncovers novel genomic regions and candidate genes of yield-related traits in upland cotton. *Theor. Appl. Genet.* **131**, 2413–2425 (2018).
88. Bates, D., Mächler, M., Bolker, B. M. & Walker, S. C. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–51 (2014).
89. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
90. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta Ct}$ method. *Methods* **25**, 402–408 (2001).
91. Senthil-Kumar, M. & Mysore, K. S. Tobacco rattle virus-based virus-induced gene silencing in *Nicotiana benthamiana*. *Nat. Protoc.* **9**, 1549–1562 (2014).

Acknowledgements

We thank X. L. Zhang, T. Z. Zhang, F. G. Li, Y. X. Zhu and Z. Jeffrey Chen for releasing sequencing and transcriptome data. We also thank the National Medium-term Gene Bank of Cotton in China and Key Laboratory for Crop Germplasm Resources of Hebei for providing cotton accessions. This work was supported by Funds of the National Key Research and Development Program to Y.Z. (2016YFD0101405) and to X.W. (2016YFD0101006), and the China Agriculture Research System (CARS15-03), the Science and Technology Support Program (16226307D) and the Top Talent Project (031601801) of Hebei Province to Z.M.

Author contributions

Z.M. conceived the research. Z.M., X.W., Y.Z. and S.T. designed the analyses. X.W., Y.Z., Z.S., S.T., Y.J. and X.Li performed genome assembly and sequencing, genomic variants and GWAS analyses. Z.M., X.W., Y.Z., Liqiang Wu, G.Z., Z.S., Z.Li, H.K., B.C., Z.Liu, Q.G., Z.W., G.W., J.Y., J.W., Y.Y., C.M., S.M., N.Wu, L.M., L.C., M.Z., A.S, Z.Y., N.Wang, Lizhu Wu, D.Z., Y.C., J.C., X.Lv, Y.L., R.S. and Y.D. performed field experiments and phenotyping. Y.Z., Z. Liu and Z.S. performed transcriptome analyses. X.W., Y.Z., Z.S., L.L. and Z.W. prepared the DNA of materials. Y.Z. and B.C. conducted gene expression and functional validation. Z.M., X.W., Y.Z. and S.T. wrote the manuscript. All authors have read and approved the paper.

Competing interests

The authors declare no competing interests.

Additional information

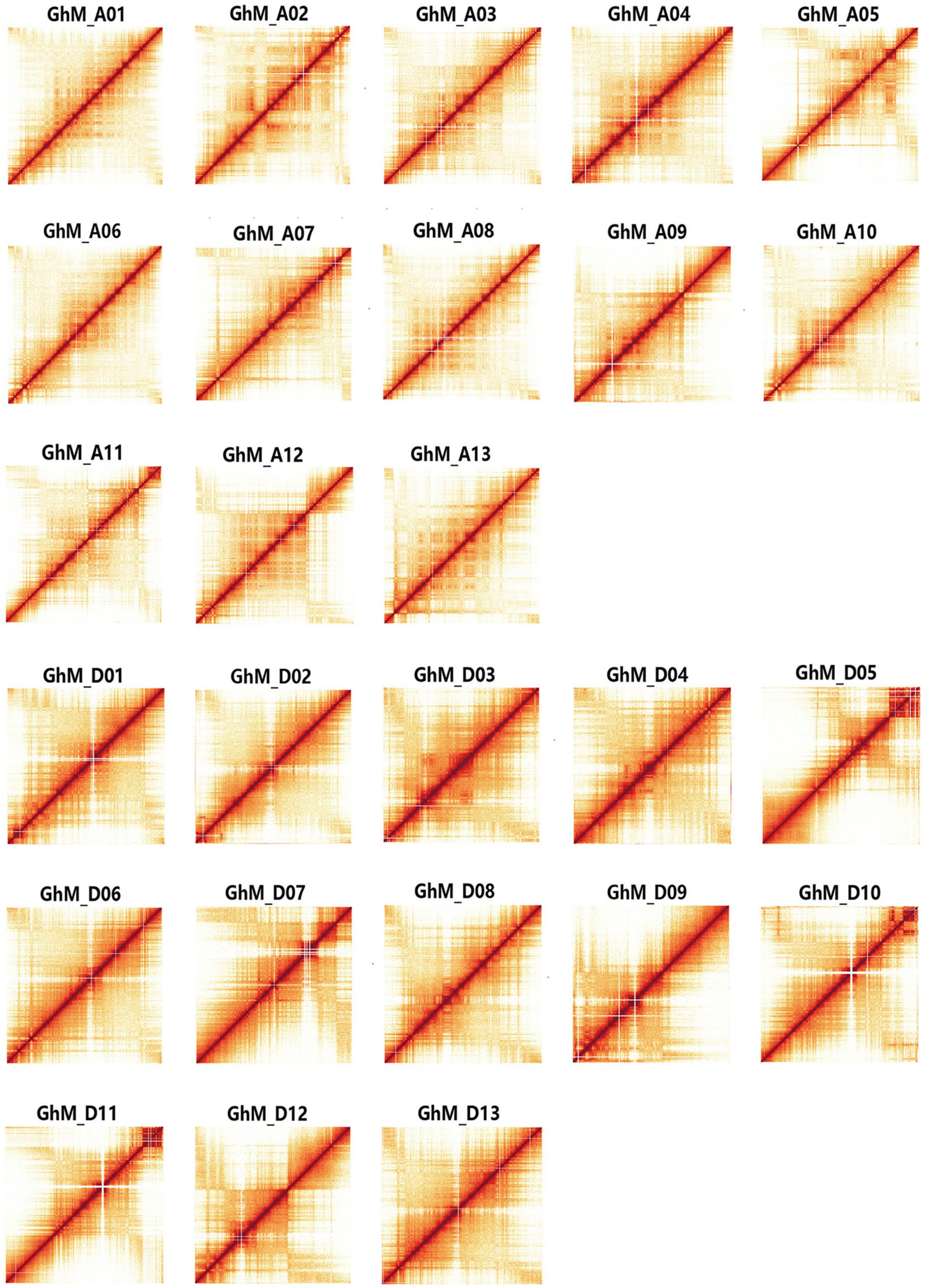
Extended data is available for this paper at https://doi.org/10.1038/s41588-021-00910-2.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00910-2.

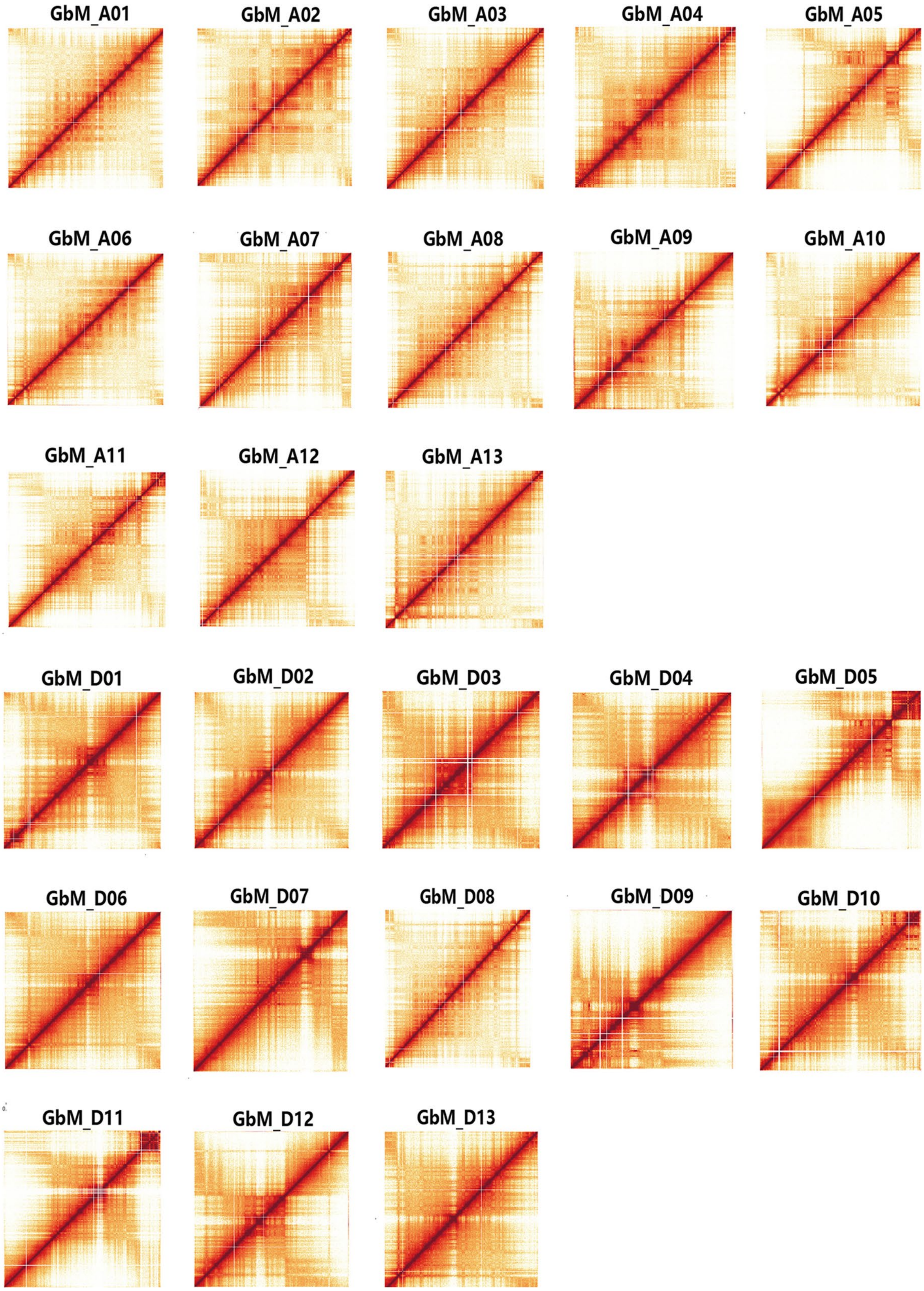
Correspondence and requests for materials should be addressed to Z.M., Y.Z., S.T. or X.W.

Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work.

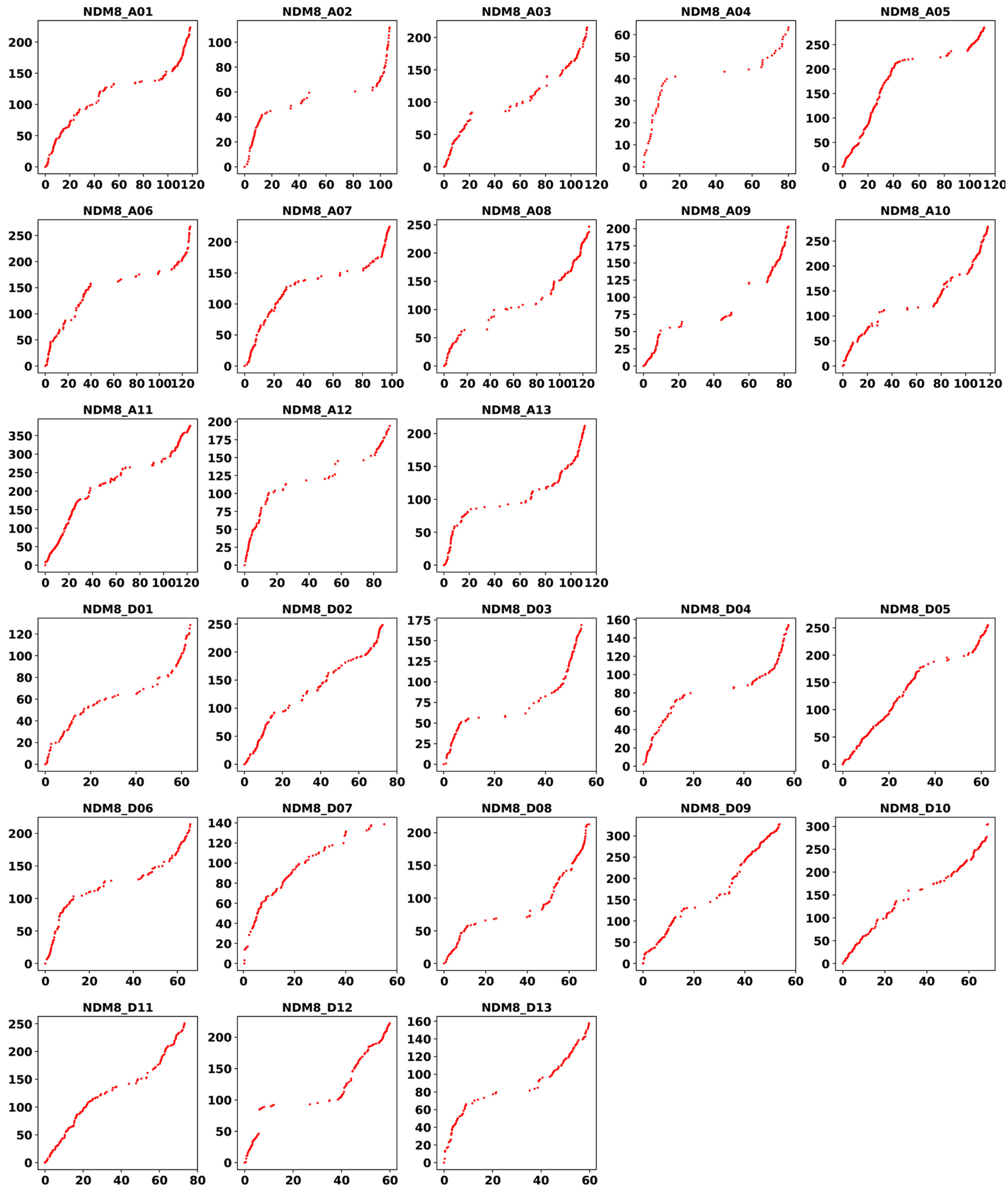
Reprints and permissions information is available at www.nature.com/reprints.



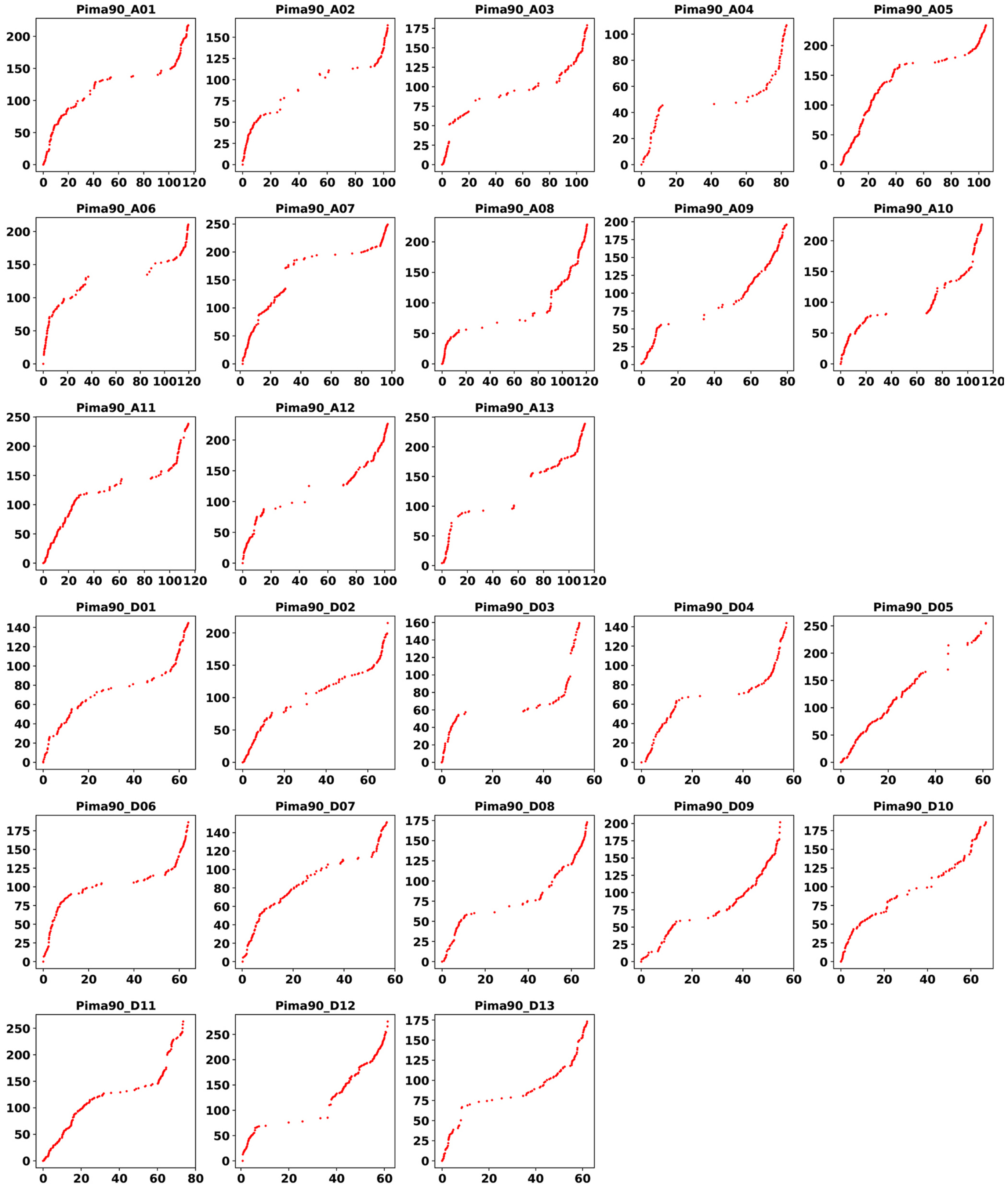
Extended Data Fig. 1 | Chromatin interactions in each chromosome of *G. hirsutum* NDM8. Each heatmap is shown at a resolution of 100 kb. The dark red dots show the high probability of interaction, and the light dots show the low probability of interaction.



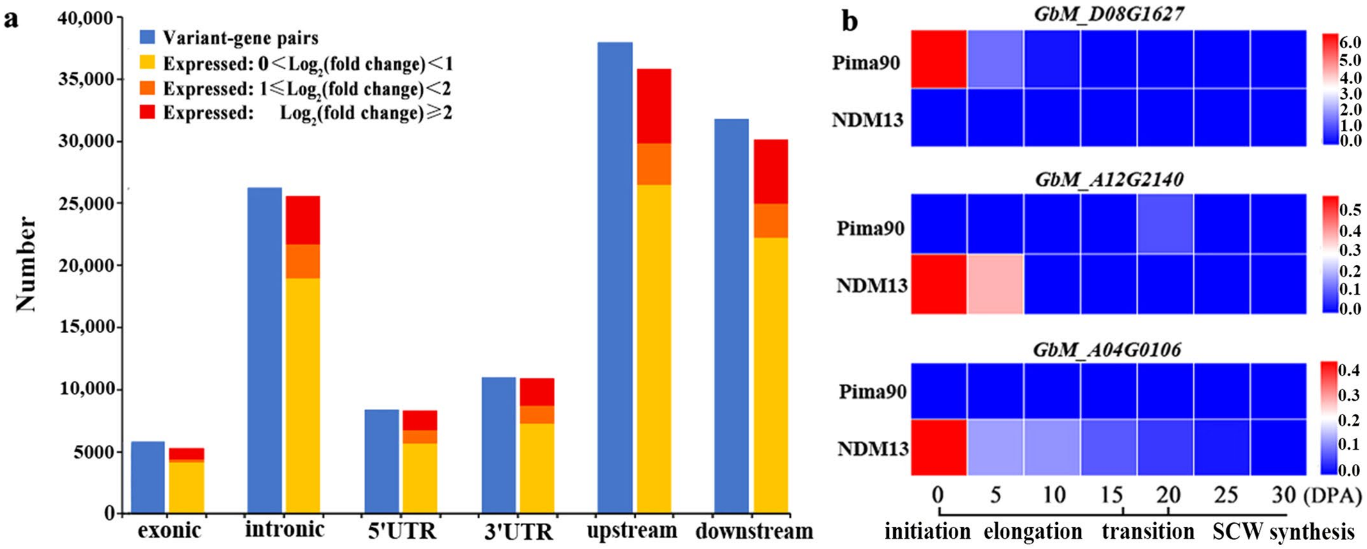
Extended Data Fig. 2 | Chromatin interactions in each chromosome of *G. barbadense* Pima90. Each heatmap is shown at a resolution of 100 kb. The dark red dots show the high probability of interaction, and the light dots show the low probability of interaction.



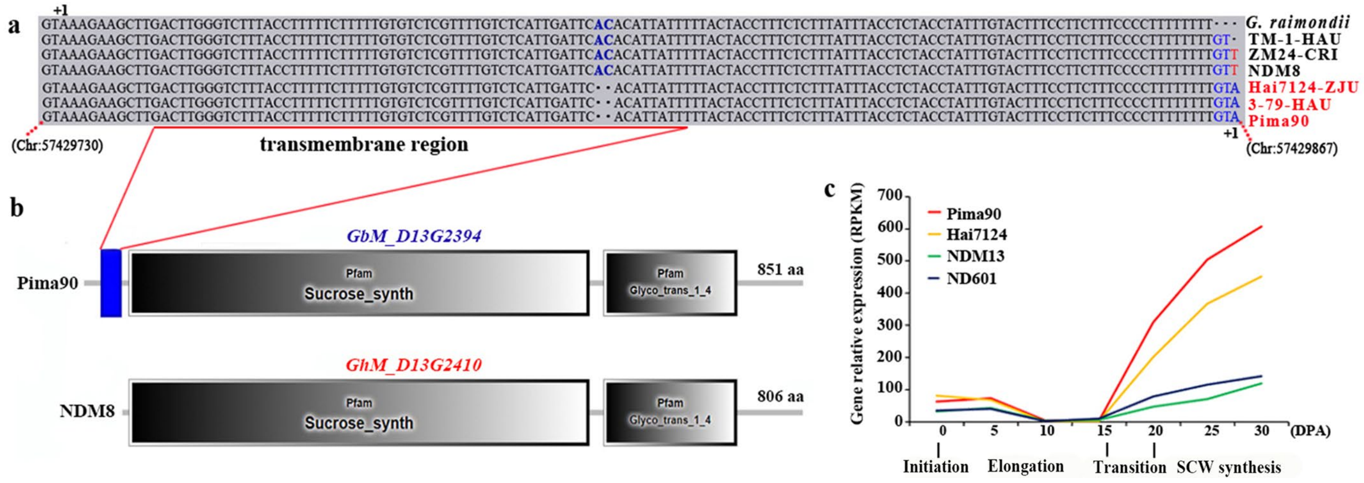
Extended Data Fig. 3 | Comparison of Hi-C directed chromosome assembly with a published genetic map between *G. hirsutum* and *G. barbadense* for each chromosome in NDM8. The x-axes represent the physical positions of the sequences (Mb) and the y-axes represent the positions of the sequences on the genetic map (cM).



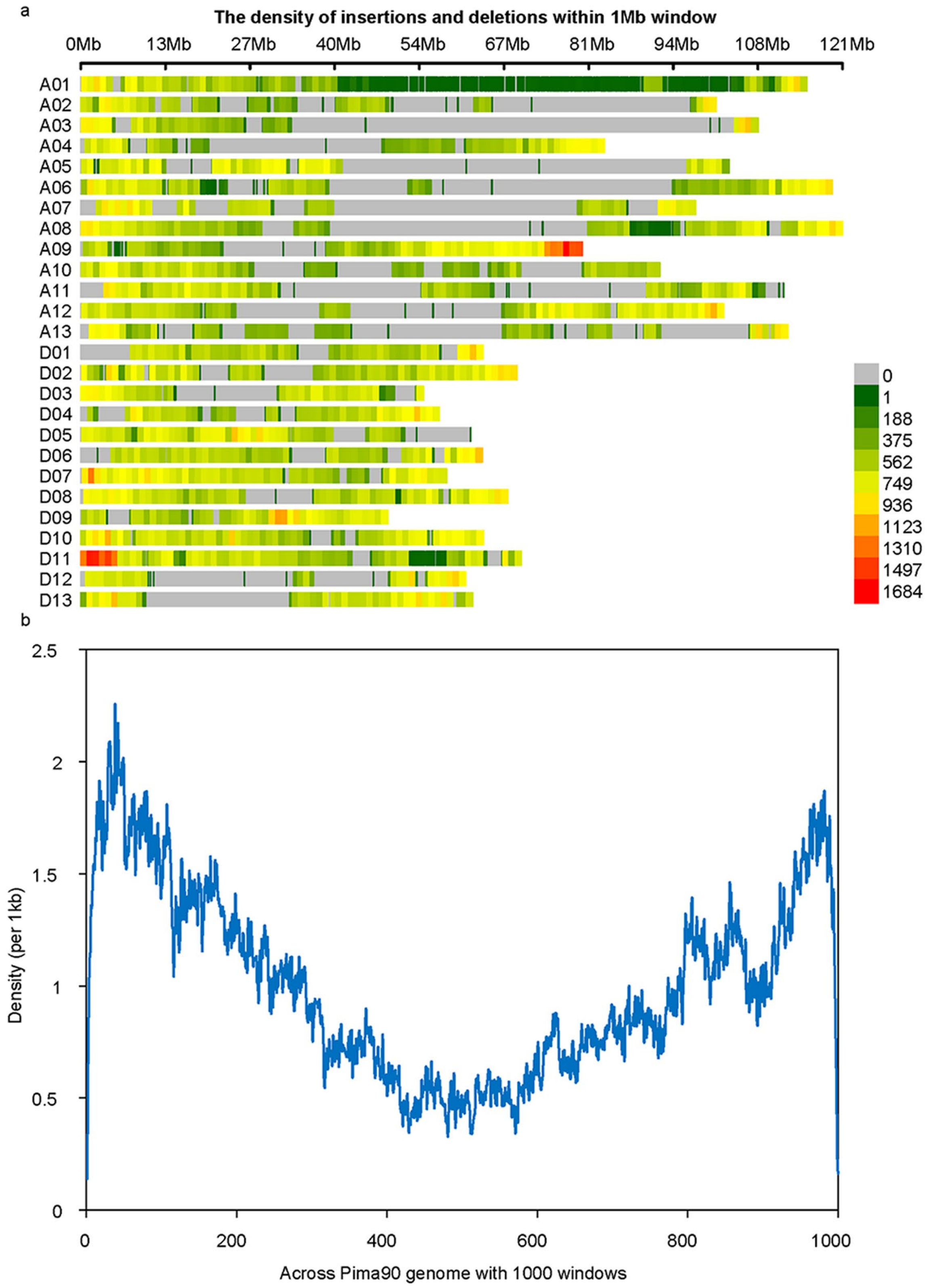
Extended Data Fig. 4 | Comparison of Hi-C directed chromosome assembly with a published genetic map between *G. hirsutum* and *G. barbadense* for each chromosome in Pima90. The x-axes represent the physical positions of the sequences (Mb) and the y-axes represent the positions of the sequences on the genetic map (cM).



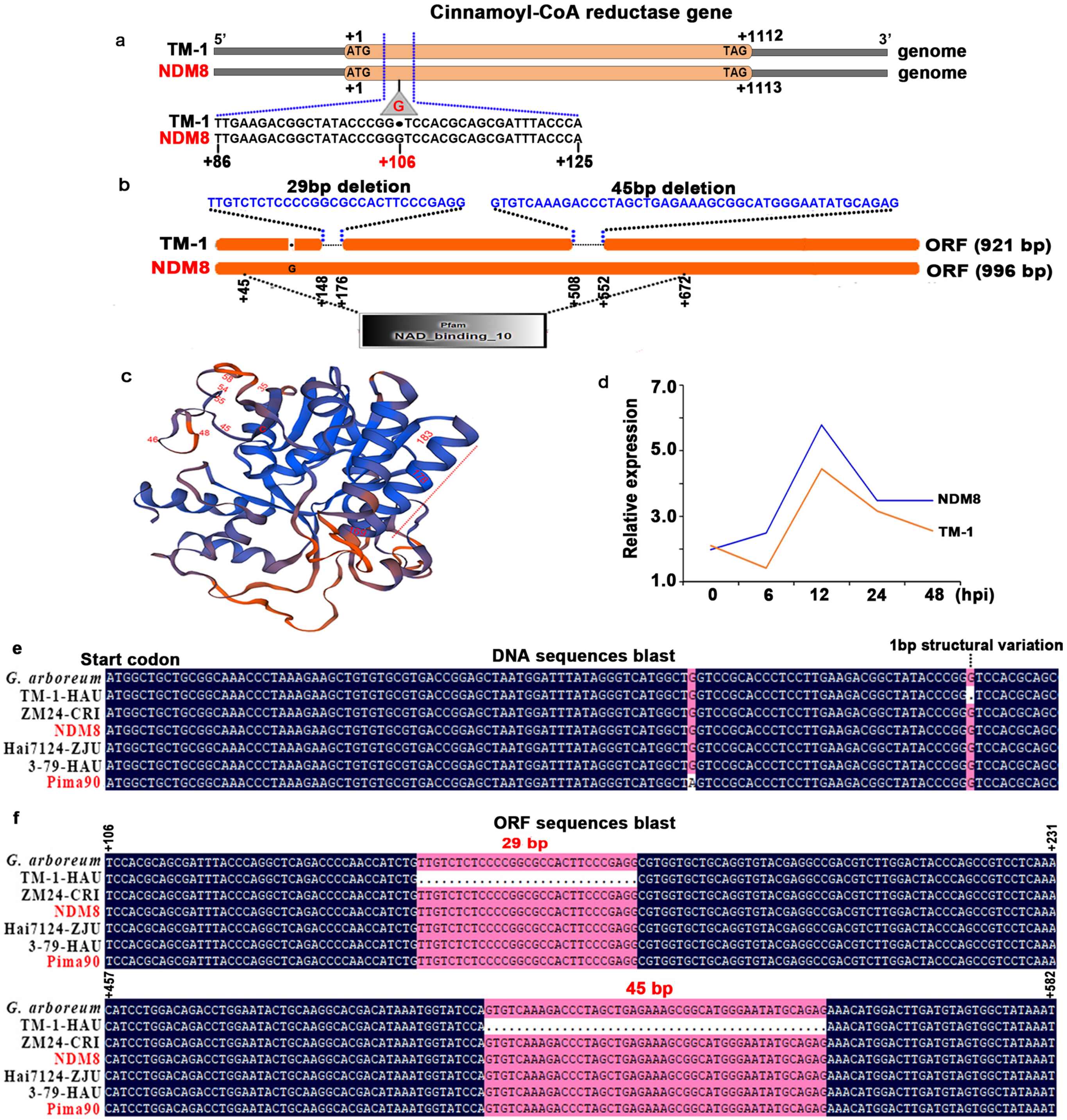
Extended Data Fig. 5 | The number of differentially expressed genes in variant-gene pairs. a, The number of differentially expressed genes with the insertion and deletion in gene and/or regulatory regions. **b**, The expression of *GbM_D08G1627*, *GbM_A12G2140* and *GbM_A04G0106*.



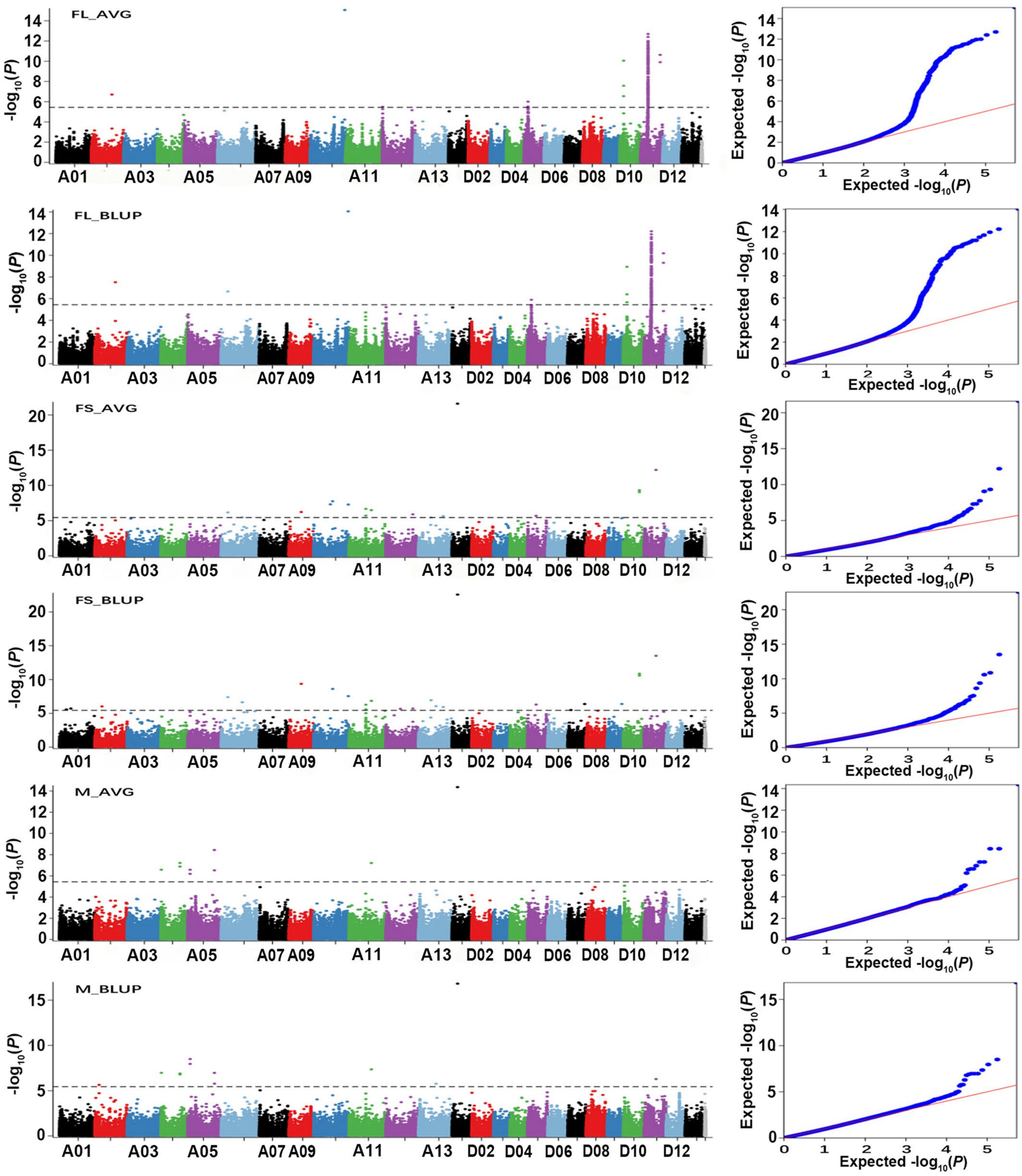
Extended Data Fig. 6 | The structure of sucrose synthase (Sus) gene in Pima90 and NDM8, and expression analysis of different stages in cotton fiber development. **a**, Comparison of *Sus* gene sequences among ancestral diploid species and cultivated tetraploid cottons. **b**, The conservative structures of the *Sus* in Pima90 and NDM8, respectively. The blue shadow rectangle indicated transmembrane region within *GbM_D13G2394*. **c**, The transcriptome of *Sus* gene in cotton varieties with different fiber quality during fiber developmental stages. The *Sus* in Pima90 with super fiber quality showed higher expression level than that in NDM8 (good fiber quality) and ND601 (common fiber quality).



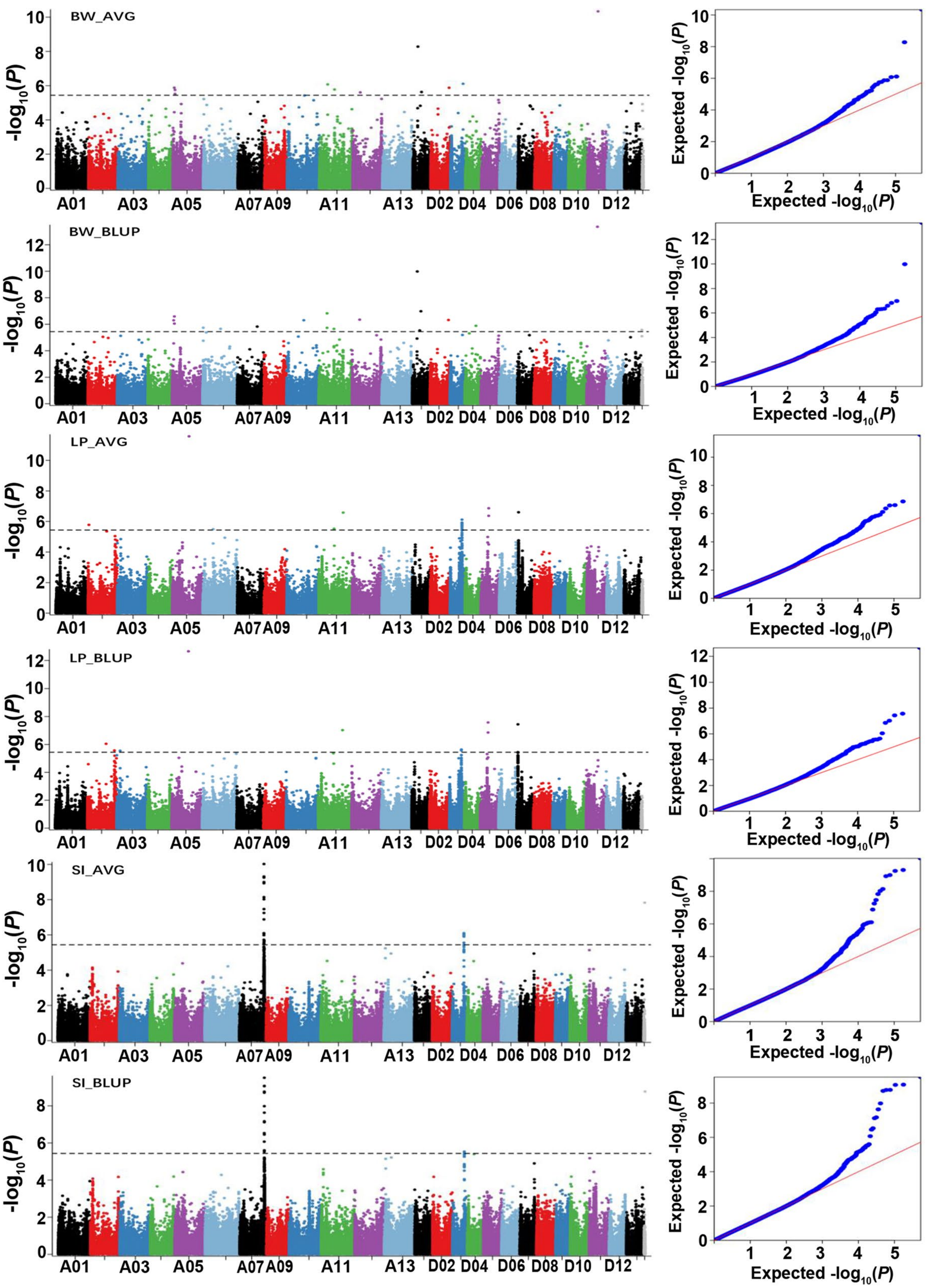
Extended Data Fig. 7 | Density distribution of insertions and deletions in Pima90. a, The density of insertions and deletions within 1 Mb window of chromosomes. **b**, The density of insertions and deletions across Pima90 genome with 1,000 windows.


Cinnamoyl-CoA reductase gene
a
TM-1
NDM8
5'
3'
+1
ATG
TAG
+1112
+1113
genome
genome
G
TM-1 TTGAAGACGGCTATACCCGG●TCCACGCAGCGATTTACCCA
NDM8 TTGAAGACGGCTATACCCGGGTCCACGCAGCGATTTACCCA
+86
+106
+125
b
29bp deletion
TTGTCTCTCCCCGGCGCCACTTCCCGAGG
45bp deletion
GTGTCAAAGACCCTAGCTGAGAAAGCGGCATGGGAATATGCAGAG
TM-1
ORF (921 bp)
NDM8
ORF (996 bp)
G
+45
+148
+176
+508
+552
+672
Pfam
NAD_binding_10
c
58
54
55
35
45
46
48
183
176
169
d
Relative expression
7.0
5.0
3.0
1.0
0
6
12
24
48 (hpi)
NDM8
TM-1
e
Start codon
DNA sequences blast
1bp structural variation
G. arboreum ATGGCTGCTGCGGCAAACCCTAAAGAAGCTGTGTGCGTGACCGGAGCTAATGGATTTATAGGGTCATGGCTGGTCCGCACCCTCCTTGAAGACGGCTATACCCGGGTCCACGCAGC
TM-1-HAU ATGGCTGCTGCGGCAAACCCTAAAGAAGCTGTGTGCGTGACCGGAGCTAATGGATTTATAGGGTCATGGCTGGTCCGCACCCTCCTTGAAGACGGCTATACCCGG.TCCACGCAGC
ZM24-CRI ATGGCTGCTGCGGCAAACCCTAAAGAAGCTGTGTGCGTGACCGGAGCTAATGGATTTATAGGGTCATGGCTGGTCCGCACCCTCCTTGAAGACGGCTATACCCGGGTCCACGCAGC
NDM8 ATGGCTGCTGCGGCAAACCCTAAAGAAGCTGTGTGCGTGACCGGAGCTAATGGATTTATAGGGTCATGGCTGGTCCGCACCCTCCTTGAAGACGGCTATACCCGGGTCCACGCAGC
Hai7124-ZJU ATGGCTGCTGCGGCAAACCCTAAAGAAGCTGTGTGCGTGACCGGAGCTAATGGATTTATAGGGTCATGGCTGGTCCGCACCCTCCTTGAAGACGGCTATACCCGGGTCCACGCAGC
3-79-HAU ATGGCTGCTGCGGCAAACCCTAAAGAAGCTGTGTGCGTGACCGGAGCTAATGGATTTATAGGGTCATGGCTGGTCCGCACCCTCCTTGAAGACGGCTATACCCGGGTCCACGCAGC
Pima90 ATGGCTGCTGCGGCAAACCCTAAAGAAGCTGTGTGCGTGACCGGAGCTAATGGATTTATAGGGTCATGGCTAGTCCGCACCCTCCTTGAAGACGGCTATACCCGGGTCCACGCAGC
f
ORF sequences blast
+106
29 bp
+231
G. arboreum TCCACGCAGCGATTTACCCAGGCTCAGACCCCAACCATCTGTTGTCTCTCCCCGGCGCCACTTCCCGAGGCGTGGTGCTGCAGGTGTACGAGGCCGACGTCTTGGACTACCCAGCCGTCCTCAAA
TM-1-HAU TCCACGCAGCGATTTACCCAGGCTCAGACCCCAACCATCTG.............................CGTGGTGCTGCAGGTGTACGAGGCCGACGTCTTGGACTACCCAGCCGTCCTCAAA
ZM24-CRI TCCACGCAGCGATTTACCCAGGCTCAGACCCCAACCATCTGTTGTCTCTCCCCGGCGCCACTTCCCGAGGCGTGGTGCTGCAGGTGTACGAGGCCGACGTCTTGGACTACCCAGCCGTCCTCAAA
NDM8 TCCACGCAGCGATTTACCCAGGCTCAGACCCCAACCATCTGTTGTCTCTCCCCGGCGCCACTTCCCGAGGCGTGGTGCTGCAGGTGTACGAGGCCGACGTCTTGGACTACCCAGCCGTCCTCAAA
Hai7124-ZJU TCCACGCAGCGATTTACCCAGGCTCAGACCCCAACCATCTGTTGTCTCTCCCCGGCGCCACTTCCCGAGGCGTGGTGCTGCAGGTGTACGAGGCCGACGTCTTGGACTACCCAGCCGTCCTCAAA
3-79-HAU TCCACGCAGCGATTTACCCAGGCTCAGACCCCAACCATCTGTTGTCTCTCCCCGGCGCCACTTCCCGAGGCGTGGTGCTGCAGGTGTACGAGGCCGACGTCTTGGACTACCCAGCCGTCCTCAAA
Pima90 TCCACGCAGCGATTTACCCAGGCTCAGACCCCAACCATCTGTTGTCTCTCCCCGGCGCCACTTCCCGAGGCGTGGTGCTGCAGGTGTACGAGGCCGACGTCTTGGACTACCCAGCCGTCCTCAAA
+457
45 bp
+582
G. arboreum CATCCTGGACAGACCTGGAATACTGCAAGGCACGACATAAATGGTATCCAGTGTCAAAGACCCTAGCTGAGAAAGCGGCATGGGAATATGCAGAGAAACATGGACTTGATGTAGTGGCTATAAAT
TM-1-HAU CATCCTGGACAGACCTGGAATACTGCAAGGCACGACATAAATGGTATCCA...AAACATGGACTTGATGTAGTGGCTATAAAT
ZM24-CRI CATCCTGGACAGACCTGGAATACTGCAAGGCACGACATAAATGGTATCCAGTGTCAAAGACCCTAGCTGAGAAAGCGGCATGGGAATATGCAGAGAAACATGGACTTGATGTAGTGGCTATAAAT
NDM8 CATCCTGGACAGACCTGGAATACTGCAAGGCACGACATAAATGGTATCCAGTGTCAAAGACCCTAGCTGAGAAAGCGGCATGGGAATATGCAGAGAAACATGGACTTGATGTAGTGGCTATAAAT
Hai7124-ZJU CATCCTGGACAGACCTGGAATACTGCAAGGCACGACATAAATGGTATCCAGTGTCAAAGACCCTAGCTGAGAAAGCGGCATGGGAATATGCAGAGAAACATGGACTTGATGTAGTGGCTATAAAT
3-79-HAU CATCCTGGACAGACCTGGAATACTGCAAGGCACGACATAAATGGTATCCAGTGTCAAAGACCCTAGCTGAGAAAGCGGCATGGGAATATGCAGAGAAACATGGACTTGATGTAGTGGCTATAAAT
Pima90 CATCCTGGACAGACCTGGAATACTGCAAGGCACGACATAAATGGTATCCAGTGTCAAAGACCCTAGCTGAGAAAGCGGCATGGGAATATGCAGAGAAACATGGACTTGATGTAGTGGCTATAAAT

Extended Data Fig. 8 | The structural variation of *CCR* gene (*GhM_A02G1731* versus *Ghir_A02G014590*). a, The location of structural variation in the genome of NDM8 against TM-1. **b**, The structural variation led to the difference in the open reading frame (ORF) between NDM8 and TM-1, and the conservative structure domain (NAD_binding_10) of *CCR* in NDM8. **c**, Three-dimensional structure of CCR (GhM_A02G1731) was obtained by homologous modeling. The second deletion (508–552) in TM-1 influenced the formation of *CCR* structure within NAD-binding domain that was indicated by red dotted line. **d**, Expression of *CCR* in resistant (NDM8) and susceptible (TM-1) cotton varieties under *V. dahliae* stress through qRT–PCR. *Ghhistone3b* was used as an internal control. **e**, Comparison of *CCR* genomic sequences among ancestral diploid species and cultivated tetraploid cottons. **f**, Comparison of *CCR* partial coding sequences among ancestral diploid species and cultivated tetraploid cottons.



Extended Data Fig. 9 | GWAS of fiber quality related traits based on accessions and structural variations. Manhattan plots and Quantile-Quantile plots using mean (AVG) and BLUP values of all environments. The genome-wide significant -log$_{10}$(P) = 5.44 is indicated by the gray dotted line. FL, fiber length; FS, fiber strength; M, micronaire value. The statistical analysis was performed with two-tailed Wald test.



Extended Data Fig. 10 | GWAS of yield related traits based on accessions and structural variations. Manhattan plots and Quantile-Quantile plots using mean (AVG) and BLUP values of all environments. The genome-wide significant $-\log_{10}(P) = 5.44$ is indicated by the gray dotted line. BW, boll weight; LP, lint percentage; SI, seed index. The statistical analysis was performed with two-tailed Wald test.

nature research

Corresponding author(s): Zhiying Ma

Last updated by author(s): May 29, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	PacBio reads were collected from single-molecule real-time (SMRT) cells on PacBio RSII and Sequel instruments; Hi-C data and paired-end reads were collected from Illumina HiSeq platform.
Data analysis	Bowtie2 (v2.3.4.3), Hiclib (20190614), FALCON assembler (v2017.11.02-16.04-py2.7), Quiver (v2.3.3), Pilon pipeline (v1.18), fragScaff (v140324), LACHESIS (v2.0), HiC-Pro (v2.10.0), BWA (v0.7.17-r1188), SAMtools (v1.9), BUSCO (v3.0), RepeatMasker (vopen-4.0.5), Tandem Repeats Finder (v4.07b), LTR FINDER (v1.07), PILER (v1.0), RepeatScout (v1.0.5), BLAST (v2.2.26), Solar (v0.9.6), GeneWise (v2.4.1), Trinity (v2.8.5), IsoSeq3 (v3.4.0), PASA (v2.3.3), Tophat (v2.1.1), Cufflinks (v2.1.1), Augustus (v3.2.3), GeneID (v1.4), GenScan (v1.0), GlimmerHMM (v3.0.4), SNAP (v2006-07-28), EvidenceModeler (v1.1.1), MUMmer (v3.2), BEDTools (v2.27), smartie-sv, Sentieon (release201808.05), SVMerge (v1.2), LUMPY (v0.2.13), BreakDancer (v1.3.6), Velvet (v1.2.10), GEMMA (v0.94.1), TreeBest (v1.9.2), OrthoMCL (V2.0.9), MCScanX (0.8), lme4 (1.1-23). Code used for the genome annotations of repetitive elements and gene structures are deposited in Zenodo DOI-minting repository with the DOI:10.5281/zenodo.4851529.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The raw sequencing data and transcriptome data of NDM8 and Pima90, and the resequencing data of 1,081 accessions are deposited in the NCBI Sequence Read Archive under the BioProject accession number PRJNA680449. The two cotton assemblies have been deposited in NCBI GenBank under the accession numbers JAHMMW000000000 and JAHMMX000000000. The versions described in this paper are version JAHMMW000000000.1 and JAHMMX000000000.1. The relevant data are also deposited in the CottonGen database https://www.cottongen.org/ (the assemblies and gene annotations) and in the website http://cotton.hebau.edu.cn/Data%20Download.html (the assemblies, gene annotations, structural variations and phenotypic data).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

1,081 cotton accessions from different origins were used for resequencing and structural variation analysis in our study, of which 1,041 samples with phenotypic data for GWAS, meeting the requirement of BLUP and average values calculation. Thirty-five cotton plants were used for VIGS assay in each independent experiment, meeting the statistical requirement.

Data exclusions

GWAS analysis was performed by the structural variations excluding those locating on At08 chromosome because large inversions existed in this chromosome, which could produce interference for the accuracy of the analysis.

Replication

In Fig. 3, qRT-PCR was conducted with two technical replicates in one experiment for each of 16 cotton varieties. VIGS assay was done twice independently. The experimental results were reliably reproduced in our study.

Randomization

The 1,081 accession genotypes were randomly planted in each location. A core collection with 401 accessions for identifying Verticillium wilt resistance were randomly planted in each environment. Two types of accessions with reference- and alternate-genotypes (each contained 8 varieties) were used for qRT-PCR in Verticillium wilt resistance test.

Blinding

The experiments were conducted blindly. All genotypes were only labeled by numbers when planting, so the investigators did not know the exact accession names.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).

Research sample

State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.

Sampling strategy

Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.

Data collection

Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.

Timing

Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.

Data exclusions

If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the

Data exclusions	*rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*
Research sample	*Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.*
Sampling strategy	*Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data collection	*Describe the data collection procedure, including who recorded the data and how.*
Timing and spatial scale	*Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Reproducibility	*Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.*
Randomization	*Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.*
Blinding	*Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Did the study involve field work? □ Yes □ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*
Access & import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*
Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Human research participants
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Antibodies

Antibodies used	*Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	*State the source of each cell line used.*
Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

□ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	*For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, gender, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☐	☐	Public health
☐	☐	National security
☐	☐	Crops and/or livestock
☐	☐	Ecosystems
☐	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☐	☐	Demonstrate how to render a vaccine ineffective
☐	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☐	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☐	☐	Increase transmissibility of a pathogen
☐	☐	Alter the host range of a pathogen
☐	☐	Enable evasion of diagnostic/detection modalities
☐	☐	Enable the weaponization of a biological agent or toxin
☐	☐	Any other potentially harmful combination of experiments and agents

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

□ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

□ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

□ All plots are contour plots with outliers or pseudocolor plots.

□ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

□ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)

Specify: functional, structural, diffusion, perfusion.

Field strength

Specify in Tesla

Sequence & imaging parameters

Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.

Area of acquisition

State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.

Diffusion MRI □ Used □ Not used

Preprocessing

Preprocessing software

Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).

Normalization

If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.

Normalization template

Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.

Noise and artifact removal

Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: □ Whole brain □ ROI-based □ Both

Statistic type for inference
(See Eklund et al. 2016)

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
□	□ Functional and/or effective connectivity
□	□ Graph analysis
□	□ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.