nature genetics

Article

https://doi.org/10.1038/s41588-025-02258-3

Genome assembly of two allotetraploid cotton germplasms reveals mechanisms of somatic embryogenesis and enables precise genome editing

Received: 21 August 2023

Accepted: 5 June 2025

Published online: 22 July 2025

Check for updates

Zhongping Xu[1,4], Guanying Wang[1,4], Xiangqian Zhu[1], Ruipeng Wang[1], Longfu Zhu[1], Lili Tu[1], Yuling Liu[2], Renhai Peng[2], Keith Lindsey[3], Maojun Wang[1] ✉, Xianlong Zhang[1] ✉ & Shuangxia Jin[1] ✉

Somatic embryogenesis is crucial for plant genetic engineering, yet the underlying mechanisms in cotton remain poorly understood. Here we present a telomere-to-telomere assembly of Jin668 and a high-quality assembly of YZ1, two highly regenerative allotetraploid cotton germplasms. The completion of the Jin668 genome enables characterization of ~30.1 Mb of centromeric regions invaded by centromeric retrotransposon of maize and *Tekay* retrotransposons, an ~8.1 Mb 5S rDNA array containing 25,190 copies and a ~75.1 Mb major 45S rDNA array with 8,131 copies. Comparative analyses of regenerative and recalcitrant genotypes reveal dynamic transcriptional patterns and chromatin accessibility during the initial regeneration process. A hierarchical gene regulatory network identifies *AGL15* as a contributor to regeneration. Additionally, we demonstrate that genetic variation affects sgRNA target sites, while the Jin668 genome assembly reduces the risk of off-target effects in CRISPR-based genome editing. Together, the complete Jin668 genome reveals the complexity of genomic regions and cotton regeneration, and improves the precision of genome editing.

Somatic cell reprogramming and regeneration demonstrate plant developmental plasticity and are essential for propagation through de novo organogenesis and somatic embryogenesis (SE)[1]. SE is also crucial for transgenic biotechnology[2], relying on coordinated cascade reactions and synergistic effects among gene networks[3]. Currently, many plant species, and even different germplasm within a species, are difficult to transform and regenerate due to genotype restriction[4,5]. Despite progress in identifying SE regulators[6] like *WIND1* (ref. 7), *BBM* and *WUS2* (ref. 8), *GRF–GIF*[9] and *TaWOX5* (ref. 4), further efforts are needed to expand transgenic receptors, enhance regeneration efficiency and shorten regeneration periods.

Allotetraploid Upland cotton (*Gossypium hirsutum* L.), an essential source of renewable textile fiber and seed oil, has undergone substantial genetic erosion due to domestication and breeding, resulting in reduced genetic diversity and the accumulation of deleterious mutations[10–12]. Genome editing offers potential solutions by introducing beneficial variations, removing harmful mutations, and enabling targeted trait improvement[13–17]. However, successful cotton regeneration is restricted to a few genotypes like Coker[18], ZM24 (ref. 11), YZ1 (ref. 19) and Jin668 (ref. 6). Among these, YZ1 and Jin668 are elite genotypes with extremely high regeneration ability[6,19], widely used in cotton research. Attempts to elucidate the molecular mechanism underlying

[1]National Key Laboratory of Crop Genetic Improvement, Hubei Hongshan Laboratory, Huazhong Agricultural University, Wuhan, China. [2]Research Base, Anyang Institute of Technology, State Key Laboratory of Cotton Biology, Anyang, China. [3]Department of Biosciences, Durham University, Durham, UK. [4]These authors contributed equally: Zhongping Xu, Guanying Wang. ✉e-mail: mjwang@mail.hzau.edu.cn; xlzhang@mail.hzau.edu.cn; jsx@mail.hzau.edu.cn

SE in cotton have identified key regulators, including *GhSPL10* for callus proliferation and embryogenic cells differentiation[20], *GhL1L1* accelerates the formation of embryonic cells[21], the *GhTCE1–GhTCEE1* module regulates cell fate determination[22], and *GhRCD1–GhMYC3–GhMYB44–GhLBD18* transcriptional cascade regulates the acquisition of somatic pluripotency[23]. Regarding epigenetic regulation, specific methylation sites have been linked to the regulation of efficient embryogenic differentiation[24]. However, none of these genes were identified in Jin668, but rather in low-regeneration genotypes, limiting their potential value for improved regeneration. While the genotype-independent shoot apical meristem (SAM) transformation system[25] has shown promise, challenges in low transformation efficiency, high chimerism and genotype-specific failures persist. However, the addition of regeneration factors can effectively mitigate these challenges. Therefore, studying the efficient regeneration mechanism in Jin668 can also promote the application of SAM transformation.

Here, we assembled a complete telomere-to-telomere (T2T) genome of Upland cotton genotype Jin668 and a high-quality genome of YZ1. Comparative genomics, chromatin accessibility and gene expression analyses shed light on the key gene regulatory networks that facilitate the ability of Jin668 to achieve regeneration. Additionally, we conducted a comprehensive evaluation of the impact of genetic variation between genotypes on the precision of CRISPR-based genome editing, emphasizing the importance of high-quality genome assembly for individual genotypes to ensure accurate genome editing for functional gene research and molecular breeding.

Results

The assembly of Jin668 and YZ1 genomes

To generate a complete genome assembly, the Upland cotton (*G. hirsutum*; $2n = 4x$ = AADD = 52) genotype Jin668 inbred line (Fig. 1a, Supplementary Fig. 1 and Extended Data Fig. 1a) was sequenced by the combination of different sequencing platforms (Supplementary Table 1). The genome was primarily assembled using 387.02 Gb of ONT reads (~169×, N50 = 52.26 kb; including ~30× reads >100 kb) and 124.88 Gb of PacBio HiFi reads (54×). The preliminary assembly was highly contiguous, consisting of 44 contigs (Supplementary Table 2), with 17 contigs representing the chromosomes. Only nine unresolved fragmented regions, likely from the most highly repetitive regions with extra-long tandem repeats. After polishing the preliminary assembly (Supplementary Table 2) and conducting assembly using Hi-C data, the remaining 23 contigs were clustered onto nine chromosomes (Supplementary Table 3), with 14 gaps persisting (Supplementary Tables 4 and 5). Subsequent gap closing, telomere patching and repeat-aware polishing using ONT, HiFi, Illumina reads and a prior HiFi-based assembly (Supplementary Table 6 and Supplementary Fig. 2) yielded a complete Jin668 genome (Table 1). With telomere repeats (AAACCCT/AGGGTTT)[26] as queries, we identified at all 52 telomeres on the termini of the 26 chromosomes (Fig. 1b and Supplementary Table 7) and validated via fluorescence in situ hybridization (FISH) assay (Supplementary Fig. 3).

Using the same workflow as the Jin668 genome assembly, but without ONT ultralong reads, the final YZ1 assembly consisted of 262 scaffolds (N50 = 108.2 Mb). The 26 pseudochromosomes accounted for 98.98% of the 2.3 Gb assembled sequences, with 45 telomeres identified across the chromosomes and only 72 gaps (Table 1, Extended Data Fig. 1b,c, Supplementary Table 8 and Supplementary Note 1).

Assembly validation and annotation

Completeness and accuracy of the Jin668 genome assembly were assessed in multiple ways. First, the unbiased genomic base distribution and no-hit aligning with microbial sequence in the nucleotide sequence database confirmed no exogenous DNA contamination (Supplementary Fig. 4 and Supplementary Table 9). Likewise, homologous alignment indicated no organelle genome sequences were integrated into the chromosome sequences (Supplementary Table 10). Genomic completeness was evaluated using BUSCO, revealing 99.2% recovery of core conserved genes (Table 1 and Supplementary Table 11). The long terminal repeat assembly index was estimated at 14.80 (Extended Data Fig. 2a), suggesting high completeness in complex repetitive regions. Additionally, the contig/chromosome (CC) ratio[27] of 1.0 indicates exceptional assembly contiguity (Table 1).

Next, we mapped all available data, including ONT, HiFi, Bionano, Illumina and Hi-C, to the assembled Jin668 genome. Among them, Illumina short reads showed an average mapping rate of 80.6% when excluding reads with secondary/supplementary alignment or low MAPQ (<10; Supplementary Table 12). The read depth of mapped sequencing from filtered high-quality long reads shows uniform coverage across all chromosomes (Fig. 2a) for either HiFi (mean coverage = 29.04, range = 27.26–34.47) or ONT (mean coverage = 14.38, range = 13.52–16.81). Mapped Hi-C (Fig. 2b and Extended Data Fig. 2b,c) and Bionano (Extended Data Fig. 2d) data showed no signs of misorientations or other large-scale structural errors. Moreover, the consensus quality values and completeness confirm the robustness of the Jin668 genome assembly (Table 1 and Extended Data Fig. 2e).

Centromeric region, characterized by massive tandem repeat monomers, were identified by aligning genotype-specific centromere-specific H3 (CenH3) chromatin immunoprecipitation sequencing (ChIP–seq) reads to Jin668 genome (Supplementary Fig. 5). Centromeric regions were identified on all 26 chromosomes, with a sharp interval of CenH3 signal on each chromosome ranging from 1.0 to 1.8 Mb. Furthermore, these centromeric regions showed consistent localization along the same chromosome in YZ1 genotype (Extended Data Fig. 1c and Supplementary Tables 13 and 14). Then, targeted validation using a purpose-built method[28] with ONT reads, indicated by a high number of single-clump *k*-mers across the centromere (Fig. 2c and Extended Data Fig. 2f), and measuring the ratio of primary to secondary variants identified by HiFi reads[29], showing a low level of secondary alleles across all centromeric sequences (Fig. 2d and Extended Data Fig. 2g), further confirmed the accuracy of the centromere assembly.

Based on these verifications, we present a complete genome of Jin668 that includes T2T assemblies for all 26 chromosomes from A_t and D_t subgenomes, comprising 2.37 Gb of nuclear DNA, with 30.1 Mb (1.4%) centromeric sequences and 1.9 Gb (82.2%) repetitive sequences (Supplementary Table 15). The long terminal repeat retrotransposons (69.8%) are mainly composed of *Gypsy* (58.8%) and *Copia* (9.3%), representing the predominant repetitive sequences (Supplementary Figs. 6 and 7). A comprehensive annotation identified a total of 74,457 protein-coding genes, with complete and accurate gene space (Table 1, Supplementary Table 11 and Supplementary Note 2). Notably, only three BUSCO genes (49273at3193, 116963at3193, and 184280at3193 in embryophyta_odb10) were missing and not fragmented, with 116963at3193 absent in Malvales from OrthoDB (v10.1). Additionally, 5S rDNA sequences with lengths of ~5.1 Mb (consisting of 15,167 repetitive monomers) and ~3.0 Mb (comprising 10,023 repetitive monomers) were identified on chromosomes A09 and D09, respectively (Supplementary Table 16). The major 45S rDNA sequences with lengths of ~37.1 Mb (4,021 repetitive monomers), ~25.8 Mb (2,799 repetitive monomers) and ~12.1 Mb (1,311 repetitive monomers) were identified on chromosomes A09, D07 and D09, respectively. However, in addition to major 45S rDNA, there are also minor 45S rDNA loci with copy numbers ranging from ~2 to ~47 on other chromosomes (Supplementary Table 16). Furthermore, FISH experiments confirmed two 5S rDNA loci and three 45S rDNA loci in Jin668 (Fig. 2e). Moreover, we further identified 116.37 Mb (5.09%) of nonredundant segmental duplications (SDs), including 41.41 Mb intrachromosomal and 85.08 Mb interchromosomal duplications (Extended Data Fig. 3, Supplementary Figs. 8 and 9 and Supplementary Note 3). The same methods were used to

https://doi.org/10.1038/s41588-025-02258-3

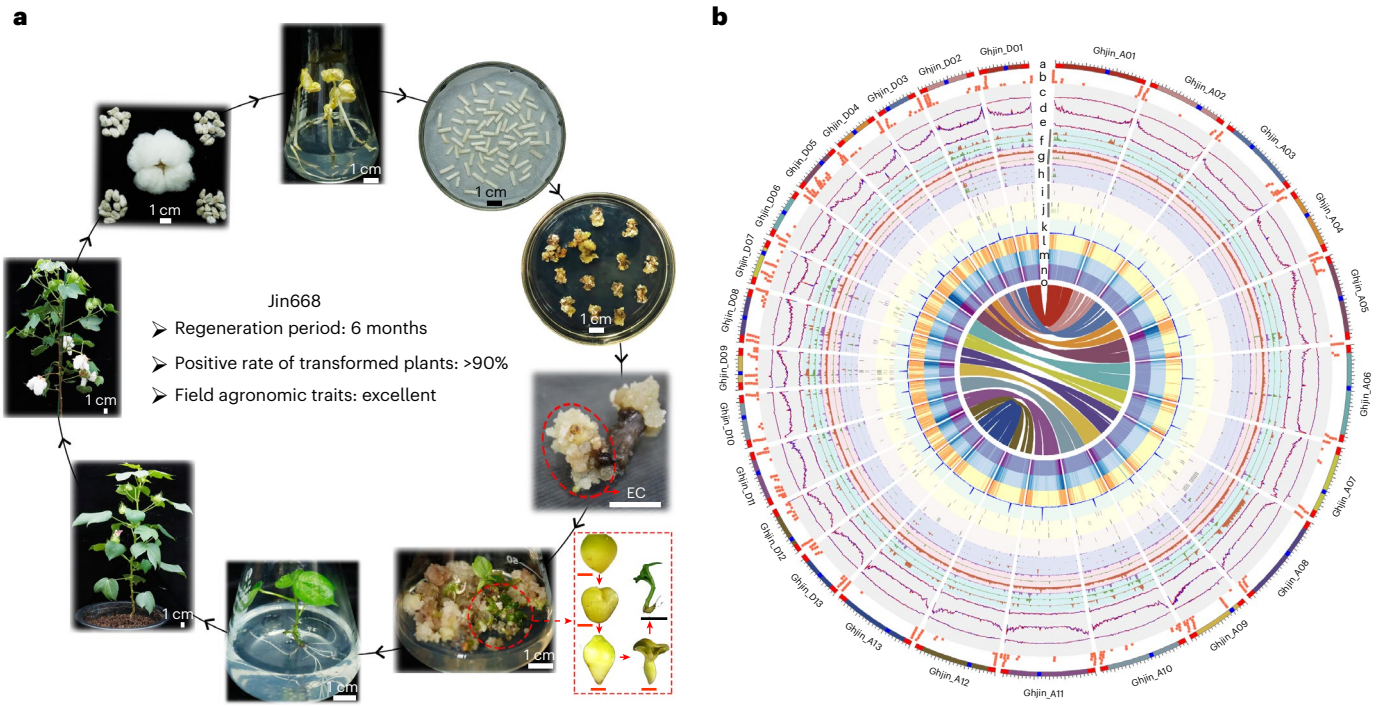


Fig. 1 | A comprehensive overview of the genetic transformation system of cotton and the complete genome features of Jin668 and YZ1. **a**, Transformation and regeneration diagram of allotetraploid cotton species. Red scale bars = 500 μm; black and white scale bars = 1 cm. **b**, Distribution of genomic features of Jin668 and YZ1. Lowercase letter a indicates chromosomes with centromeres (blue) and telomeres (red). Lowercase letter b indicates satellite repeats in Jin668. Lowercase letters c–e indicate the GC content, gene and repeat density in Jin668 (blue lines) and YZ1 (red lines). Lowercase letters f–j indicate SNPs, InDels, PAVs, inversions and translocation density between Jin668 with TM-1, YZ1 and ZM24. Lowercase letter k indicates CENH3 (log_2(ChIP/input)). Lowercase letter l indicates 5mC DNA methylation levels. Lowercase letters m and n indicate histone modification levels of H3K4me3 and H3K27ac. Lowercase letter o indicates syntenic blocks between the homoeologous A and D chromosomes. The densities in plots in c–e are represented in 1 Mb with overlapping 200-kb sliding windows. For k–n, all tracks show data in Jin668. EC, embryogenic callus.

demonstrate and annotate a high-quality assembly of the YZ1 genome (Supplementary Note 4, Supplementary Figs. 10 and 11 and Supplementary Table 17).

Global view of centromere architecture

The assembled Jin668 genome provided insights into the architecture of centromeres. First, chromosome-scale profiles of mean mappability values within 10-kb windows showed a precipitous decrease in mappability at specific regions, consistent with CENH3 log_2(ChIP/input) signals (Supplementary Fig. 12). These centromeric regions were characterized by the multimegabase tandem repeat with high sequence identity and uniform orientation (Supplementary Fig. 13). Additionally, centromeric *Ty3/Gypsy*-like retroelements (CRG1, JQ009328; CRG2, JQ009329) and pericentromeric retrotransposons (GhCR1–GhCR4, KF517432–KF517435) were all identified in the centromeric regions of Jin668 genome, except for Ghjin_D08 (Supplementary Tables 18 and 19). Interestingly, Ghjin_D08 contains 194-bp monomers (Fig. 3a) with high sequence similarity to previous reports[30] and has evolved four variants with similar sequences identity ranged from 42.4% to 99.5% (Fig. 3b). The FISH assay further confirmed their exclusively colocalization with the primary constriction of a pair of homologous chromosomes (Fig. 3c,d and Supplementary Fig. 14a). The abundance statistics showed that these 194 bp monomers were repeated an astonishing ~3,543 times within the centromeric region of Ghjin_D08, accumulating to a total length of 691 kb (Fig. 3a). Interestingly, the 194 bp monomers formed high-order repeats (HOR), in addition to being located in the centromeric region, were also found in the pericentromeric regions, with ~3,384 copies totaling 682 kb in length. Moreover, the pericentromeric HOR was located approximately 4 Mb away from the centromeric HOR. Furthermore, combining the FISH experiments and genomic analysis, we speculated that the 194 bp HOR spans both the centromeric and pericentromeric regions of Ghjin_D08 in Jin668 (Fig. 3c,d and Supplementary Fig. 14b).

Within the centromeres of Jin668, we identified 2,298 Class I retrotransposons and 136 Class II DNA transposons (Supplementary Table 20). Further annotation of transposon-protein-domains using reverse transcriptase and integrase sequences revealed a high abundance of centromeric retrotransposon of maize (*CRM*; A plant centromere-specific *Ty3/Gypsy* lineages long terminal repeat retrotransposons[31]), followed by *Tekay* lineages, rather than *ATHILA* that is found in *Arabidopsis*[32]. The intact *Tekay* has a mean length of 11.87 kb, longer than *CRM* (10.20 kb; Supplementary Fig. 15). Additionally, the historical dynamics analysis showed that *CRM* elements were relatively younger than *Tekay* (Fig. 3e), and centromeric *CRM* were significantly younger than those outside centromeres (91.43%, $P = 4.64 \times 10^{-146}$; Fig. 3f). Interestingly, *Tekay* and *CRM* also showed distinct historical dynamics between intrachromosomal and interchromosomal comparisons, as well as between A_t and D_t subgenomes (Supplementary Note 5 and Supplementary Figs. 16–19). Beyond transposable elements (TEs), we also identified 70 centromere-contained genes with potential contributions to basic biological functions (Supplementary Table 21).

To assess genetic and epigenetic features of the centromeres, we defined centromere midpoints using maximum CENH3 ChIP–seq enrichment and analyzed chromosome arms along telomere–centromere axes using a proportional scale (Extended Data Fig. 4). As expected, CENH3 signals were highly enriched in proximity to centromeres, which were relatively GC-rich compared to the AT-rich at chromosome arms. The density of genes, *Copia* and MuLE-MuDR drops when approaching the centromeres, while *Gypsy* density increased. Gene density was also closely correlated with H3K4me3 and H3K27ac

https://doi.org/10.1038/s41588-025-02258-3

Table 1 | Genome assembly and annotation statistics for two highly regenerable *Gossypium hirsutum* L. germplasms, Jin668 and YZ1

	Jin668	YZ1	ZM24	TM-1-HAU[10]	TM-1-WHU[47]	TM-1-UTX[48]	TM-1 v.3.1 (ref. 49)
Sequencing platform	Nanorpore UL+HiFi	HiFi+nanorpore	PacBio Sequel	PacBio RS II	PacBio	PacBio Sequel+ PacBio RS II	PacBio Sequel II
Genome sequencing depth (×)	169+54	18+19	54	90	81.6	94.06	116.73
Genome sequencing depth BioNano (>150 kb)	220	119	–	92	–	–	–
Genome sequencing depth (×) of Hi-C	104	104	–	–	–	40	172
Estimated genome size (Gb)	2.3						
Assembly quality evaluation	–	–	–	–	–	–	–
Assembly levels	T2T	Platinum	Platinum	Platinum	Platinum	Platinum	Platinum
Number of contigs	26	344	3,717	4,790	1,235	6,733	314
Number of scaffolds	26	75	2,247	2,190	342	1,025	249
Sequenced genome size (Gb)	2.37	2.31	2.31	2.35	2.29	2.3	2.28
Number of contigs>1Mb (%)	100.0% (26)	17.2% (59)	18.4% (685)	12.3% (587)	42.02% (519)	7.96% (536)	27.1% (85)
Contig N90 (Mb)	64.87	19.31	0.42	0.23	1.34	0.17	13.91
Contig N50 (Mb)	109.28	58.95	1.98	1.9	5.02	0.78	39.95
Number of scaffolds>1Mb (%)	100.0% (26)	36% (27)	1.2% (26)	1.2% (27)	7.60% (26)	2.53% (26)	10.4% (26)
Scaffold N90 (Mb)	64.87	60.68	53.05	56.43	57.94	59.44	59.07
Scaffold N50 (Mb)	109.28	108.21	93.25	97.78	106.04	108.14	106.53
Longest scaffold (Mb)	127.67	128.08	121.9	124.06	128.49	128.19	126.93
GC content (%)	35.08	34.43	34.4	33.42	34.37	34.36	34.32
Repetitive sequences (%)	82.22	82.16	72.1	69.86	64.06	73.21	70.8
Annotated protein-coding genes	74,457	73,959	73,707	70,199	74,350	75,376	75,663
BUSCO completeness of assembly (%)	99.2	99.0	99.1	99	99.2	99	99.5
BUSCO completeness of annotation (%)	99.9	99.7	99.6	99.4	97.8	99.5	98.6
Number of gaps	0	72	1,472	2,566	893	5708	65
CC ratio	1.00	13.23	142.96	184.23	47.50	258.96	12.08
Merqury consensus quality value	45.87	44.98	–	–	–	–	–
Merqury completeness	98.76	99.18	–	–	–	–	–
Number of telomeres	52	45	15	19	17	5	41

modifications, both of which decreased towards centromeres, consistent with their roles in promoting transcriptional activation[33]. Furthermore, using HiFi reads, we observed dense CpG DNA methylation and transformation of three-dimensional genomic architecture across centromeres (Supplementary Note 5).

Genomic basis underlying regeneration via SE

A comparative genome analysis of Jin668 and YZ1, which exhibit higher regeneration ability, alongside ZM24 with medium regeneration ability[11] and TM-1 (a genetic and genomic standard line for Upland cotton with nonregeneration ability), was carried out to explore the molecular mechanisms underlying somatic cell regeneration. The A_t subgenomes have a higher proportion of inversions and translocations than the D_t subgenomes, especially on chromosomes A1–A5 and A10 (Fig. 4a and Extended Data Fig. 5a). Among annotated genes from genotypes with regeneration (Jin668, YZ1 and ZM24) and nonregeneration (TM-1) ability, after ruled out the influence of gene annotation differences and removing orthologous genes present in TM-1 from three regenerable genotypes (Supplementary Note 6), approximately 560 genes were found to be specifically present among all three regenerable genotypes. These genes were enriched in hormone-related processes, such as auxin polar transport and auxin response (*PIN* and *ARFs*), positive regulation of cytokinesis and auxin homeostasis (Supplementary Table 22).

The alignment of genome sequences revealed the presence of more syntenic blocks (average 99.0% versus 96.0%) within the regeneration genotype compared to the nonregeneration genotype (Fig. 4a and Supplementary Tables 23 and 24). The remaining genomic variations between Jin668, YZ1, ZM24 and TM-1 mainly included nearly 1.08 million single nucleotide polymorphisms (SNPs), 0.24 million insertions and deletions (InDels), 37 inversions (220 bp to 18.11 Mb), 151 intratranslocations (260 bp to 735.91 kb) and 560 presence/absence variations (PAVs) on average, with the most pronounced variations on chromosome A08 (Fig. 4b, Extended Data Fig. 5b, Supplementary Table 25–27 and Supplementary Note 6). When investigating the correlation between the genomic variation and the 853 SE-related genes derived from a previous

https://doi.org/10.1038/s41588-025-02258-3

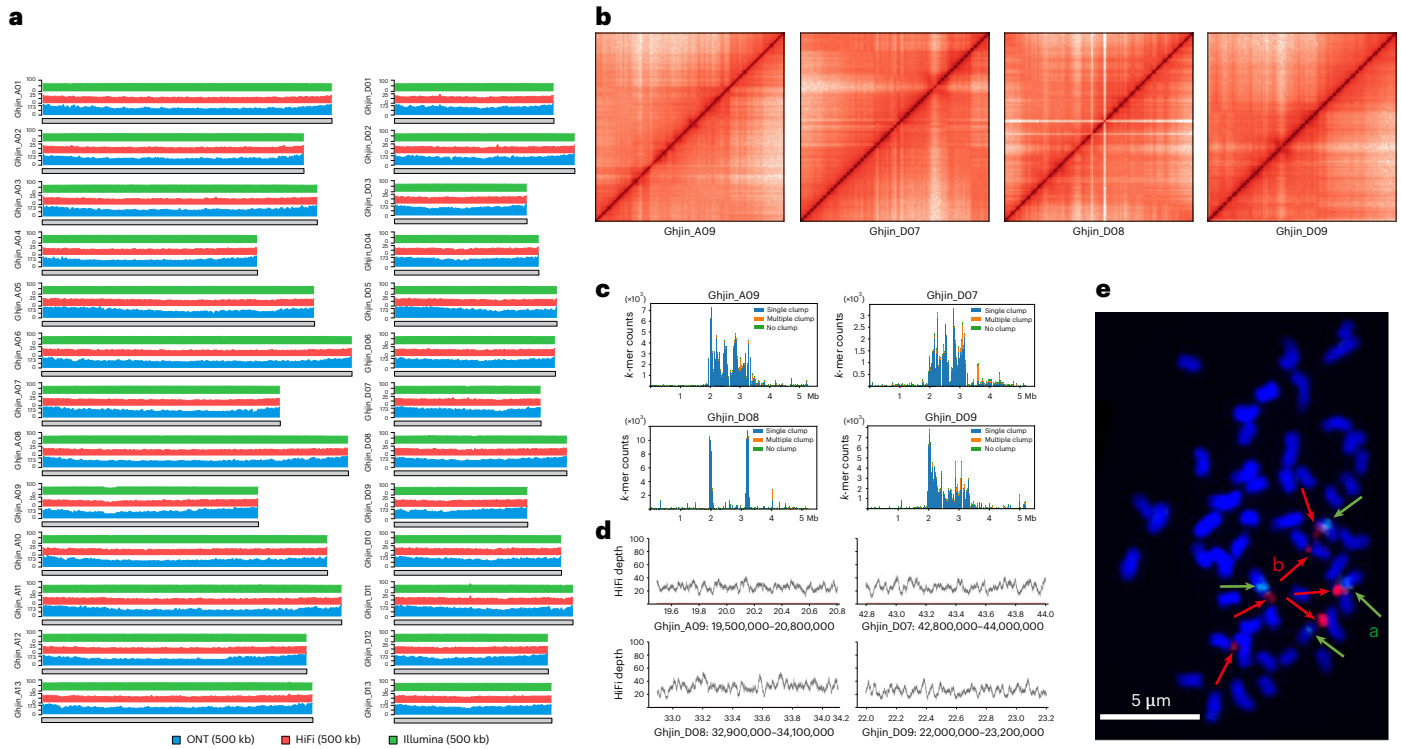


Fig. 2 | Completeness and accuracy validation of both chromosome arms and centromeres in Jin668 genome. a, Uniform whole-genome coverage of mapped Illumina, HiFi and ONT reads in Jin668. **b**, Hi-C map of the Jin668 genome showing chromosome-wide all-by-all interactions. **c**, The distribution of different types of unique *k*-mers along the centromeric regions in chromosomes of Jin668. Each bar shows the number of different types of *k*-mers in a bin of length 20 kb. The blue bars represent single-clump *k*-mers, which suggest a good base-level quality. While the orange (multiple clumps) and green (no clumps) bars suggest a low base-level quality in the region. **d**, NucFreq plot of centromeric regions in chromosomes of Jin668. HiFi coverage depth (black) along with secondary allele frequency (red) for all centromeres and surrounding regions. **e**, FISH assay using the Jin668 5S (green arrow; a) rDNA and 45S (red arrow; b) rDNA as probes in the same metaphase cells. The 5S rDNAs were located on chromosomes A09 and D09. The 45S rDNAs were located on chromosomes A09, D07 and D09. Experiment was repeated thrice independently with similar results. Scale bar = 5 μm. Here only chromosomes Ghjin_A09, Ghjin_D07, Ghjin_D08 and Ghjin_D09 are depicted for **b**–**d**, while the remaining chromosomes are depicted in Extended Data Fig. 2.

single-cell transcriptome study[34] (Supplementary Table 28 and Supplementary Note 6), certain regions of variation exhibited a positive correlation with SE-related genes (permTest, $P < 0.05$; Extended Data Fig. 5c,d). These regions may represent mutation hotspots influencing regeneration. Among them, a total of 378 and 556 gene body and promoter regions (2 kb upstream of TSS (transcription start site)) exhibited mutations distinguishing regenerative and nonregenerative genotypes. We also used TM-1, Jin668, YZ1 and ZM24 as reference genomes for single nucleotide variant (SNV) and InDel detection across nine additional genotypes (Supplementary Table 29) with varying regeneration ability[6]. Obviously, TM-1 has a high proportion of gene variations, indicating a relatively distant evolutionary divergence (Fig. 4c). Conversely, Jin668, YZ1 and ZM24 have a relatively close genetic relationship, consistent with their high regeneration ability. Moreover, the relatively stable nucleotide diversity (π) within each population suggested consistent genetic diversity. This consistency implies that the observed genetic variations were not predominantly driven by global changes in allele frequency associated with geographical origin (Supplementary Note 7 and Supplementary Fig. 20a).

To investigate the potential role of TEs in cotton regeneration, we compared TE distributions within SE-related genes in the TM-1, Jin668, YZ1 and ZM24 genomes. TM-1 exhibited a relatively high proportion of TE insertions, particularly in promoter regions compared to gene bodies (2,088 versus 663 in TM-1, 1,291 versus 350 in Jin668, 1,272 versus 261 in YZ1 and 1,396 versus 394 in ZM24), with Mutator-like elements being the predominant type (Supplementary Table 30). Notably, 32 promoter regions (27 regeneration specific and 5 nonregeneration specific) and 45 coding regions (17 regeneration specific and 28 nonregeneration specific) of orthologous genes containing TE insertion between high and nonregeneration ability genotypes (Supplementary Fig. 20b). Examples include *LEC*, a key regulator of plant embryogenesis[35]; *WOX4*, which promotes procambial development[36]; and auxin transport, response and synthesis genes (*AUX1*, *ARF6*, *YUC10* and *LBD*) among others (Supplementary Table 31). As expected, differences in TE insertion patterns correlated with variations in gene expression during SE (Fig. 4d).

Dynamic chromatin and transcriptome profiles in initial SE

During SE, a hierarchical gene regulatory network[37], including epigenetic modifications that affect the activity of transcription factors (TFs), which in turn regulate the expression of target genes to influence endogenous auxin and cytokinin levels, ultimately inducing cell reprogramming and embryo development[3]. To document chromatin accessibility dynamics and transcriptional responses during SE, we performed ATAC–seq and RNA-seq on hypocotyl explants from high regeneration (Jin668 and YZ1) and nonregeneration (TM-1 and ZB1092) genotypes (Supplementary Note 8 and Supplementary Fig. 21a) at 0, 0.5, 1, 6 and 12 h after induction (HAI; Supplementary Note 9, Extended Data Fig. 6a and Supplementary Fig. 21b). After stringent quality control (Supplementary Note 10 and Extended Data Fig. 6b–f), joint *k*-means clustering ($k = 5$) revealed progressive chromatin accessibility gains in regenerative genotypes, whereas nonregenerative genotypes exhibited accessibility losses (Fig. 5a and Supplementary Fig. 22).

During the initial stage of callus induction, expression of embryonic cell marker genes, such as *LECs*, *AGL15*, *SERK1* and *LBDs*, was upregulated in Jin668, but was not detected in TM-1. By 6–12 HAI, Jin668

https://doi.org/10.1038/s41588-025-02258-3

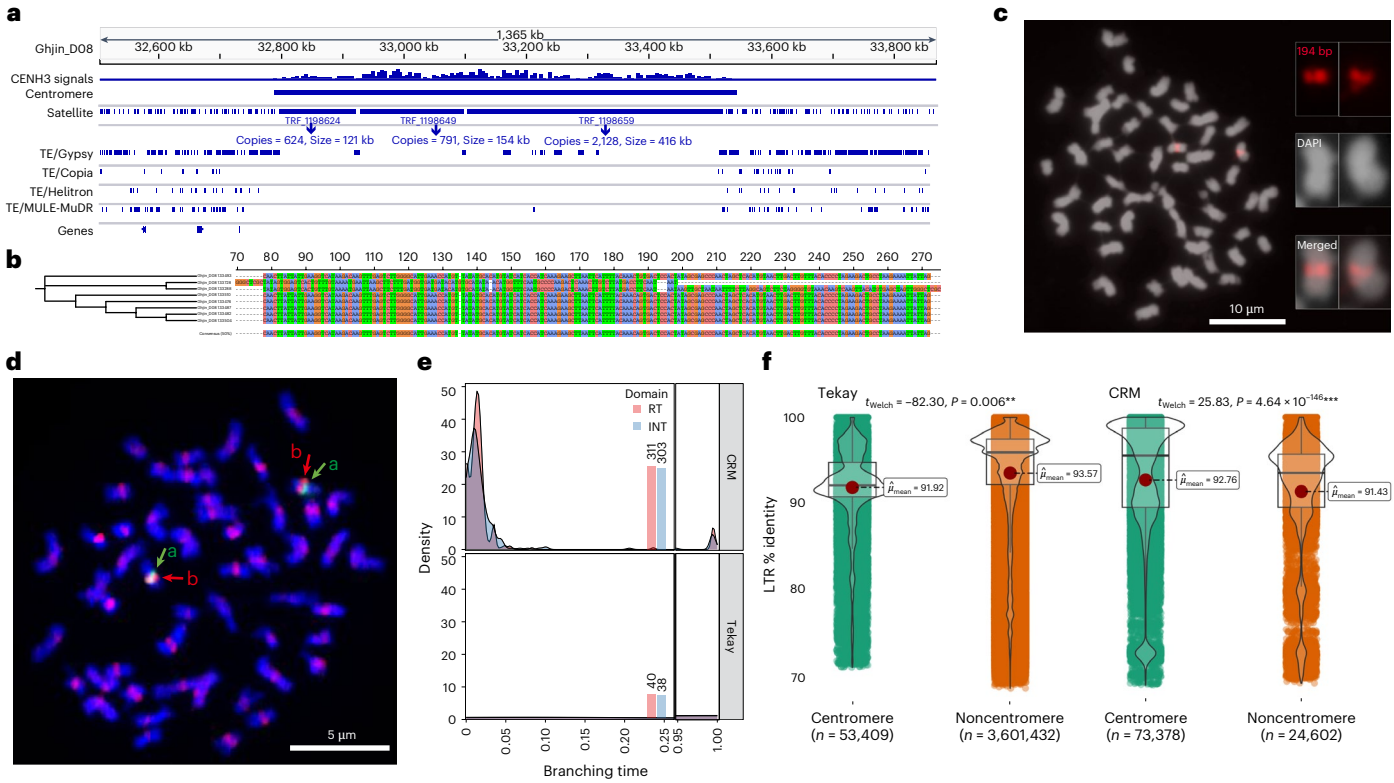


Fig. 3 | Centromeric characteristics of Jin668. a, The number of repetitions and cumulative length of 194 bp satellite repeat in the centromere region of Ghjin_D08 chromosome. **b**, Comparison of sequence identity of the Jin668 194 bp satellite repeat library. **c**, FISH assay using the Jin668 194 bp repeats (red) as probes was performed on metaphase cells, revealing their colocalization with the primary constriction. Experiment was repeated independently with similar results. Scale bar = 10 μm. **d**, FISH assay using the Jin668 194 bp satellite repeats (green arrow; a) and Gr344 retrotransposon[50] (red arrow; b) as probes in the same metaphase cells. Experiment was repeated independently with similar results. Scale bar = 5 μm. **e**, Average age of *CRM* and *Tekay* lineages in the centromeric regions was revealed through RT and INT protein domains. The bar plot columns on the right give the number of RT and INT domains present in the Jin668 genome. **f**, Comparison of sequence identity between the centromere and chromosome arm regions of *CRM* and *Tekay* lineages. For box plots, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers. Differences between two groups were evaluated by two-sided Welch's *t*-test, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. RT, reverse transcriptase; INT, integrase.

maintained high expression of the auxin-response-related gene *IAA*, auxin efflux carrier gene *PIN*, callus-inducing-related gene *WOX13*, and embryogenesis-related gene *AGL15*. In contrast, TM-1 failed to sustain the expression of these genes (Fig. 5a, Supplementary Fig. 23 and Supplementary Note 11), indicating an inability to establish auxin polarity and maintain downstream auxin responses, likely hindering SE initiation.

We further curated 688 nonredundant TF binding motifs, incorporating external SE-related TFs to accurately infer TF dynamics at gene promoters within chromatin accessibility regions (Fig. 5b). During the initial stage of callus induction, Jin668 showed specific enrichment of morphogenetic regulatory TFs binding motifs, including *WUS*, *WOX*, *LEC2*, *RAP2.6*, *EIL* and *ERF*, which were absent in TM-1. After 6 HAI, the auxin polarity regulator *KAN2* (ref. 38) was highly enriched in Jin668 (Supplementary Fig. 24), indicating that an enhanced capacity for the polar transport of auxin is a characteristic of this regenerative genotype. Moreover, after 12 HAI, TFs involved in morphogenesis regulation, auxin polarity and meristem maintenance—such as *WOX11*, *AP2* and *RIN*—were significantly enriched in Jin668 but not TM-1 (Fig. 5b, Supplementary Figs. 24 and 25 and Supplementary Note 11).

Overall, these results suggest that dynamic chromatin accessibility changes during SE in Jin668 activate the transcription of key regulatory genes in response to external hormone stimulation. This cascade influences the homeostasis of endogenous hormones and the establishment of polar auxin transport. The accumulation and establishment of endogenous auxin gradients is expected to promote dedifferentiation and the formation of embryogenic callus (Fig. 5c).

Next, we curated SE-related genes from analyses of the Jin668, YZ1 and TM-1 genomes, alongside epigenetic and transcriptome data, and used their expression levels at all induction time points to construct the SE-related co-expression network (Fig. 5d). Obviously, this network clearly comprises five modules, with the majority of enriched genes participating in wound induction, auxin response, ethylene response and biosynthesis, epigenetic regulation and hormonal signaling. There is also a large number of TFs involved in the regulation of each module, including *AP2*/*ERF*, *bHLH*, *NAC*, *WRKY* and *MYB*. Notably, in the epigenetic regulatory module, DNA-directed RNA polymerase (DdRP) exhibited strong interactions with the auxin transporter *PIN*, *ERF* and morphogenesis factor *AGL*. Within the auxin response module, *SAUR* and other genes linked to auxin, cytokinin, ethylene, and growth-regulating factor exhibited a strong regulatory relationship. Additionally, genes encoding *MYC2*, *bHLH*, serine/threonine phosphatase regulators associated with ethylene and auxin signaling were identified in the hormone response module. Further tracking the genomic locations of genes involved in these SE-related regulation networks found that some were located within centromere and SD regions (Supplementary Table 32). Thus, the identification of SE-related regulation networks in Jin668 facilitates further exploration of the regulatory dynamics underlying SE.

AGL15 instead of *AUX1* is essential for SE-based regeneration

Jin668 and TM-1 have a similar regeneration process and both can induce callus formation. Notably, Jin668 callus can further differentiate into loose, bright yellow embryonic callus, whereas TM-1 callus

https://doi.org/10.1038/s41588-025-02258-3

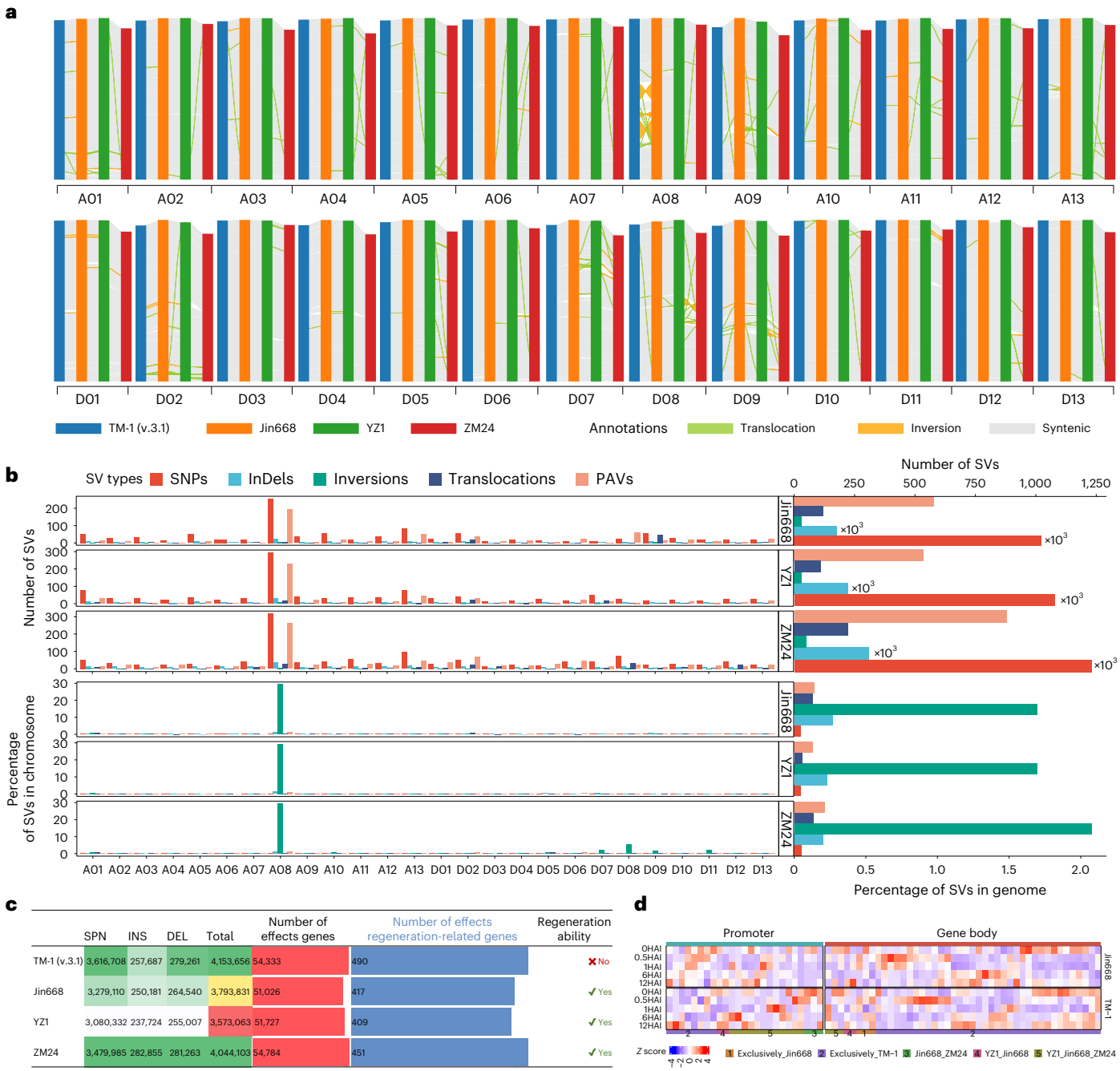


Fig. 4 | In-depth comparative analysis of the Jin668 versus TM-1, YZ1 and ZM24. a, Structural rearrangements between the four allotetraploids of Upland cotton, as well as ancestral genome of each chromosome. **b**, Statistical variation between Jin668 versus TM-1, YZ1 and ZM24 genomes. **c**, Statistical variation for nine resequencing samples when Jin668, TM-1, YZ1 and ZM24 were used as reference genomes. **d**, The heatmap illustrates the impact of TE insertion polymorphism in both gene and promoter regions of SE-related genes between Jin668 and TM-1, on their respective expression during SE.

remains compact, globular and green colored nonembryogenic callus (Fig. 6a). To address the genetic mechanisms underlying SE-based regeneration (Extended Data Fig. 7), we identified candidate genes specific to regenerative genotypes (Jin668, YZ1 and ZM24), especially for Jin668-specific, by analyzing genomic mutations, TE polymorphisms (Supplementary Table 33), and dynamic expression patterns during SE in Jin668. Notably, the auxin influx carrier gene *AUX1* and the morphogenetic gene *AGL15* caught our attention because of persistent upregulation in regenerative genotypes and modestly low expression in nonregenerative genotypes (Extended Data Fig. 8a). To further validate their functions in SE-based regeneration, we constructed CRISPR/Cas9 vectors targeting these genes and performed genetic transformations using Jin668, YZ1 and TM-1 as receptors. Compared with the control line (P7N), the hypocotyls of the *AGL15* and *AUX1* mutants in both Jin668, YZ1 and TM-1 showed a reduced somatic embryo formation rate, and the callus proliferation rate (CPR) of *AGL15* and *AUX1* mutant was also significantly reduced in YZ1 at 20 days post-induction ($P < 0.05$; Fig. 6b). Correspondingly, the overexpression of *AGL15* led to a significant increase in the CPR in both YZ1 and TM-1. However, the overexpression of *AUX1* did not result in a significant increase in CPR in TM-1. Notably, neither *AGL15* nor *AUX1* overexpression significantly altered the CPR of Jin668, possibly due to its inherently high capacity for differentiation. Following further induction and at 60 days post-induction, it was observed that all genotypes displayed a significant increase in CPR

https://doi.org/10.1038/s41588-025-02258-3

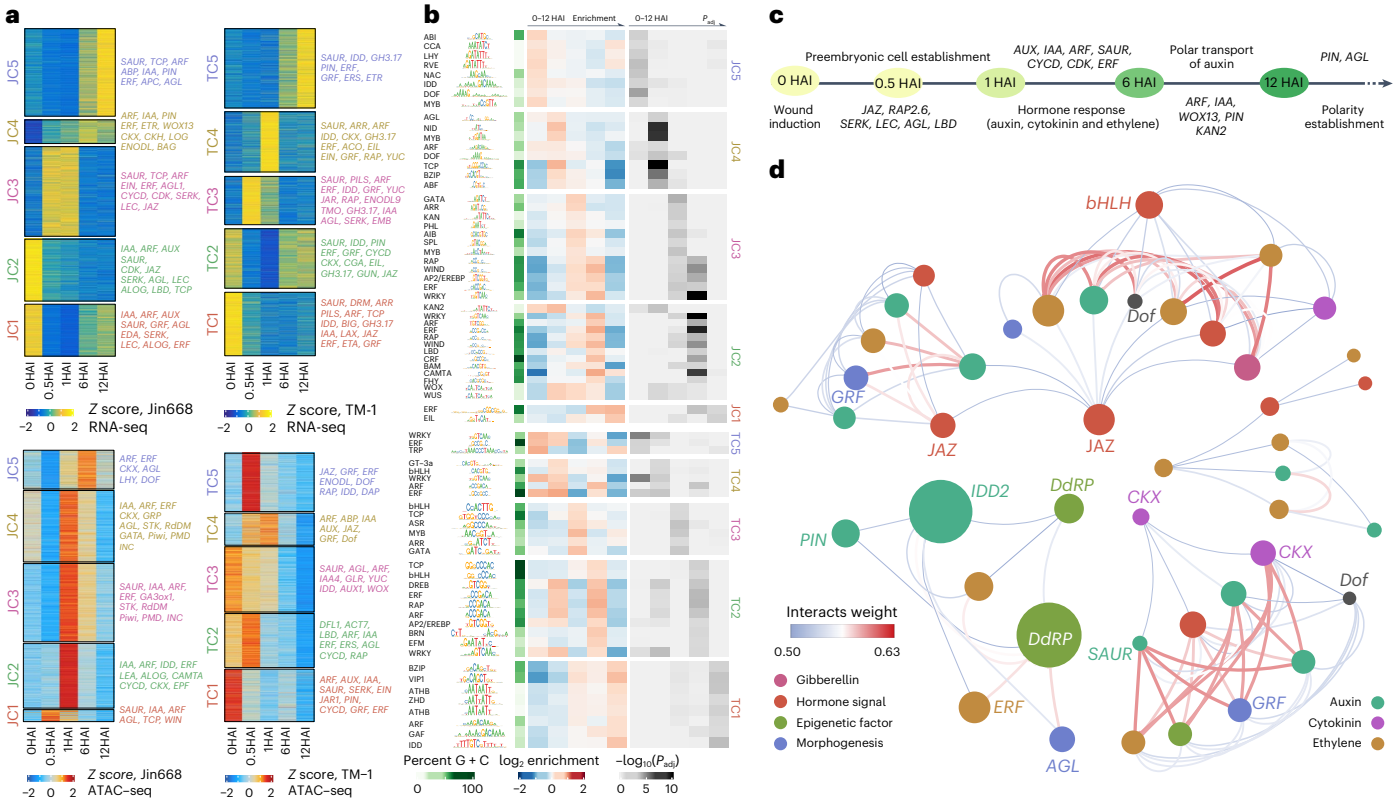


Fig. 5 | Overview of chromatin accessibility and transcriptome dynamics in Jin668 and TM-1 during SE. a, Heatmap of differentially expressed genes and differentially accessible peaks sorted by *k*-means clustering across the samples collected at different time points (0–12 HAI). Color bar, accessibility *z* score of differentially expressed genes identified by RNA-seq and differentially accessible peaks identified by ATAC–seq. The representative genes are shown on the right. **b**, HOMER DNA-motif enrichment analyses of accessible peaks in Jin668 (top) and TM-1 (bottom). The significantly (FDR < 0.05) enriched binding motifs of TF families were selected and shown here. All significant enrichment TF families are shown in Supplementary Figs. 24 and 25. **c**, The major physiological responses and their corresponding core regulatory genes during SE in Jin668. **d**, The regulatory network of SE-related genes. The known regulatory genes, classified based on their similarity with the homologous sequences in *Arabidopsis*, were labeled with different colors. The size of the nodes in the network represents the level of closeness degree. The thickness and color of the edges between the nodes indicate the interaction weight values between the two nodes. The thicker edge in red color indicates the interaction relationship with higher feasibility.

after the overexpression of *AGL15* ($P < 0.05$; Extended Data Fig. 8b). Moreover, the paraffin sectioning of hypocotyls obtained from *AGL15* and *AUX1* mutants demonstrated a pronounced reduction in cell proliferation within the vascular tissues, while overexpression lines displayed an increased proliferation, particularly in response to *AGL15* overexpression (Fig. 6c and Extended Data Fig. 8c). Collectively, these results preliminarily indicate that *AGL15* has a crucial role in SE-based regeneration within specific genetic backgrounds, as reflected by comparisons of Jin668 and YZ1 with TM-1 genotypes.

Jin668 genome improves the accuracy of cotton genome editing

In previous genome research, the lack of a Jin668 reference genome required using TM-1 genome for single-guide RNA (sgRNA) design[15]. However, considerable genetic variations between TM-1 and Jin668 (Fig. 4) pose challenges for accurate sgRNA design. To address this, we systematically investigated the impact of genetic variants in Jin668, ZM24 and TM-1 on sgRNA design. This involved using CRISPOR with standard parameters to screen the sgRNAs and potential off-targets of CRISPR/Cas9 (PAM (protospacer adjacent motif), 5′-NGG-3′) and CRISPR/Cas12a (Cpf1; PAM, 5′-TTTN-3′) for all protein-coding genes (Supplementary Fig. 26a). For each genome sequence, there are similar fractions of bases for individual genomes, while total numbers of genomic and exonic PAMs were different (Fig. 7a), indicating that genomic variants indeed alter PAM recognition. Subsequently, using CRISPOR, more than six million (6.4–6.6 M) and three million (3.2–3.4 M) high-quality sgRNA targets of Cas9 and Cas12a were identified, with notable fluctuations across genomes. Likewise, the number of inferior sgRNA targets also showed disorder fluctuations (Fig. 7a and Supplementary Fig. 26b), reflecting the influence of genetic variation on target site prediction.

To eliminate discrepancies in sgRNA on-target and off-target assessments due to gene number variations, we identified 1:1:1 orthologous genes across these genomes (Supplementary Table 34). Notably, the SNPs and InDels variants in Jin668 and TM-1 orthologs distinctly influenced the number of corresponding sgRNA on-target and off-target sites (Fig. 7b). Surprisingly, the number of on- and off-target sites of high-quality Cas9 and Cas12a sgRNAs identified for orthologous genes showed random and irregular distributions. Moreover, the benefit of assembling the Jin668 genome is that we could identify more precise sgRNAs and new potential off-targeting sites (Fig. 7c), aiding in the elimination of suboptimal sgRNAs. Furthermore, compared with Cas12a, Cas9 displayed a higher number of potential off-target sites (Fig. 7c), indicating a clear way to improve the editing accuracy of Cas9. Further analysis of orthologous genes with 94.7% sequence identity confirmed that using Jin668 as a reference genome enhances the accuracy of sgRNA design (Fig. 7d and Supplementary Note 12).

To experimentally ascertain the impact of genomic variation on the precision of sgRNA editing, we selected a sgRNA target present in Jin668 but absent in TM-1 for genetic transformation using hypocotyl from Jin668 and TM-1 as the receptors (Extended Data Fig. 9a). After three months of cultivation on the culture medium, barcode

https://doi.org/10.1038/s41588-025-02258-3

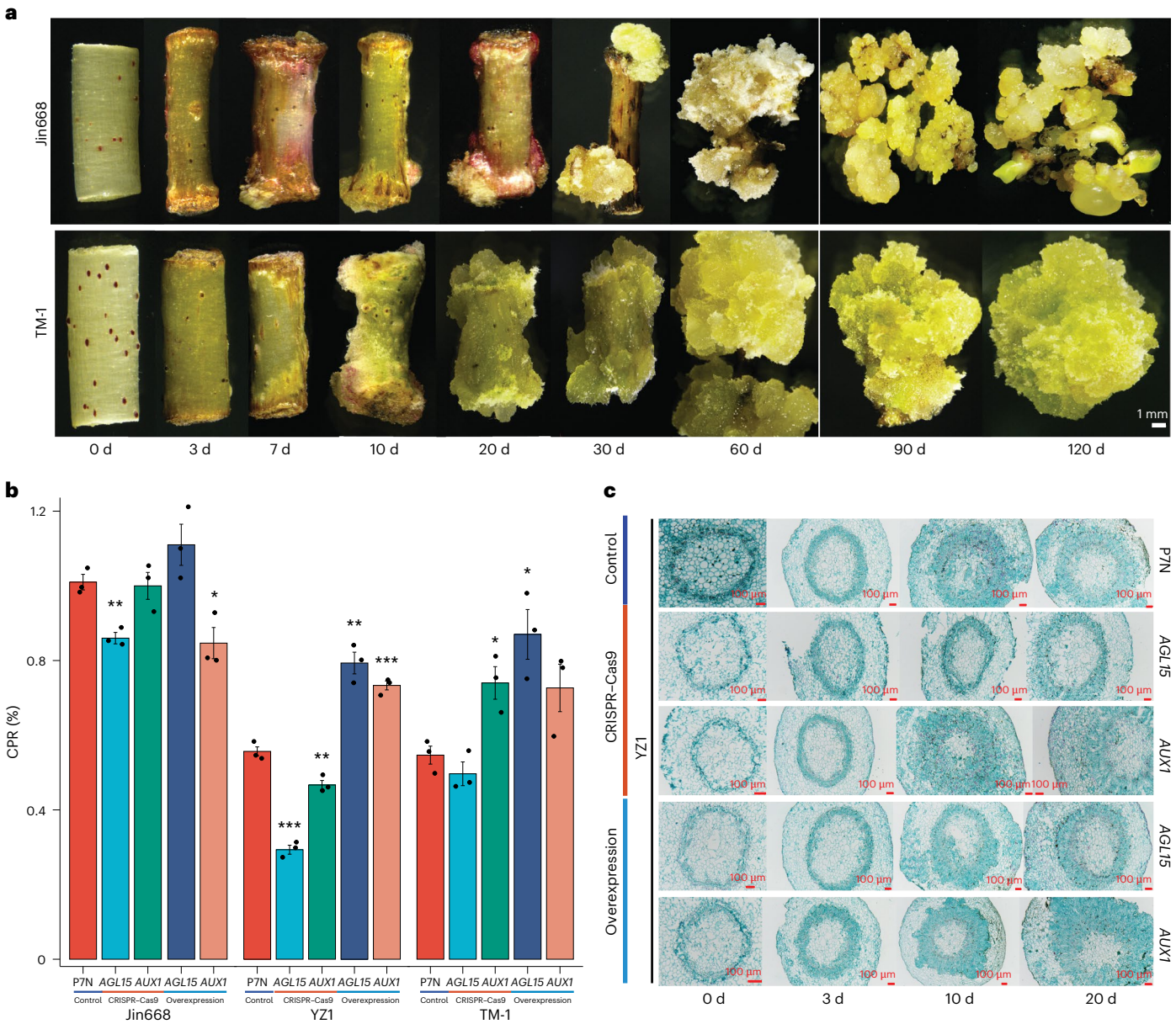


Fig. 6 | Functional verification of regeneration-related genes. a, The SE-based regeneration process and corresponding time points (unit, day) of Jin668 and TM-1. Scale bars = 1 mm. **b**, The CPR of CRISPR edited, overexpression and control lines at 20 days postinduction. Data are represented as mean ± s.d. Differences between groups were evaluated by two-sided Student's *t*-test. $n = 3$ independent biological replicates. The P7N was used as control. Statistical significance is indicated as follows: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. **c**, The paraffin sections of CRISPR edited, overexpression and control lines at different days postinduction. Scale bars = 100 μm.

sequencing[39] of callus DNA revealed up to 23.6% of the sequenced reads have mutations in Jin668, while no effective mutations were detected in TM-1 (Extended Data Fig. 9b,c and Supplementary Table 35).

To support genome editing experiments and functional genomics research, we established T2TCotton-Hub (Supplementary Fig. 27), providing genomic resources, CRISPR tools, gene expression data and interactive analysis features (Supplementary Note 13, Extended Data Fig. 10 and Supplementary Figs. 28–31).

Discussion

Plant genetic transformation introduces beneficial genes into target genotype, while CRISPR-based gene editing generates genetic variation for traits improvement[40], de novo domestication, and redomestication of wild species[17,41]. However, the indirect regeneration mode of cotton via SE hinders progress in research and breeding. Although the genotype-independent SAM transformation in cotton offers an alternative[25], yet practical challenges remain due to low efficiency, high chimerism (particularly when using CRISPR systems with inherently low efficiency), and genotype-specific barriers. In contrast, the efficient hypocotyl-based system remains an ideal approach. Moreover, research on transformable genotypes can identify regeneration factors applicable to SAM transformation. Here, leveraging advances in sequencing and genome assembly algorithms, we generated a T2T genome assembly of allotetraploid cotton genotype Jin668, resolving complex centromeric, telomeric and rDNA regions and providing valuable genomic resources for investigating the exceptional regeneration ability of Jin668.

Using the genome sequences, with 'transcriptome selection' model[42] and hierarchical regulatory network[37], we implemented a comprehensive strategy to identify SE regulators. Key assumptions

https://doi.org/10.1038/s41588-025-02258-3

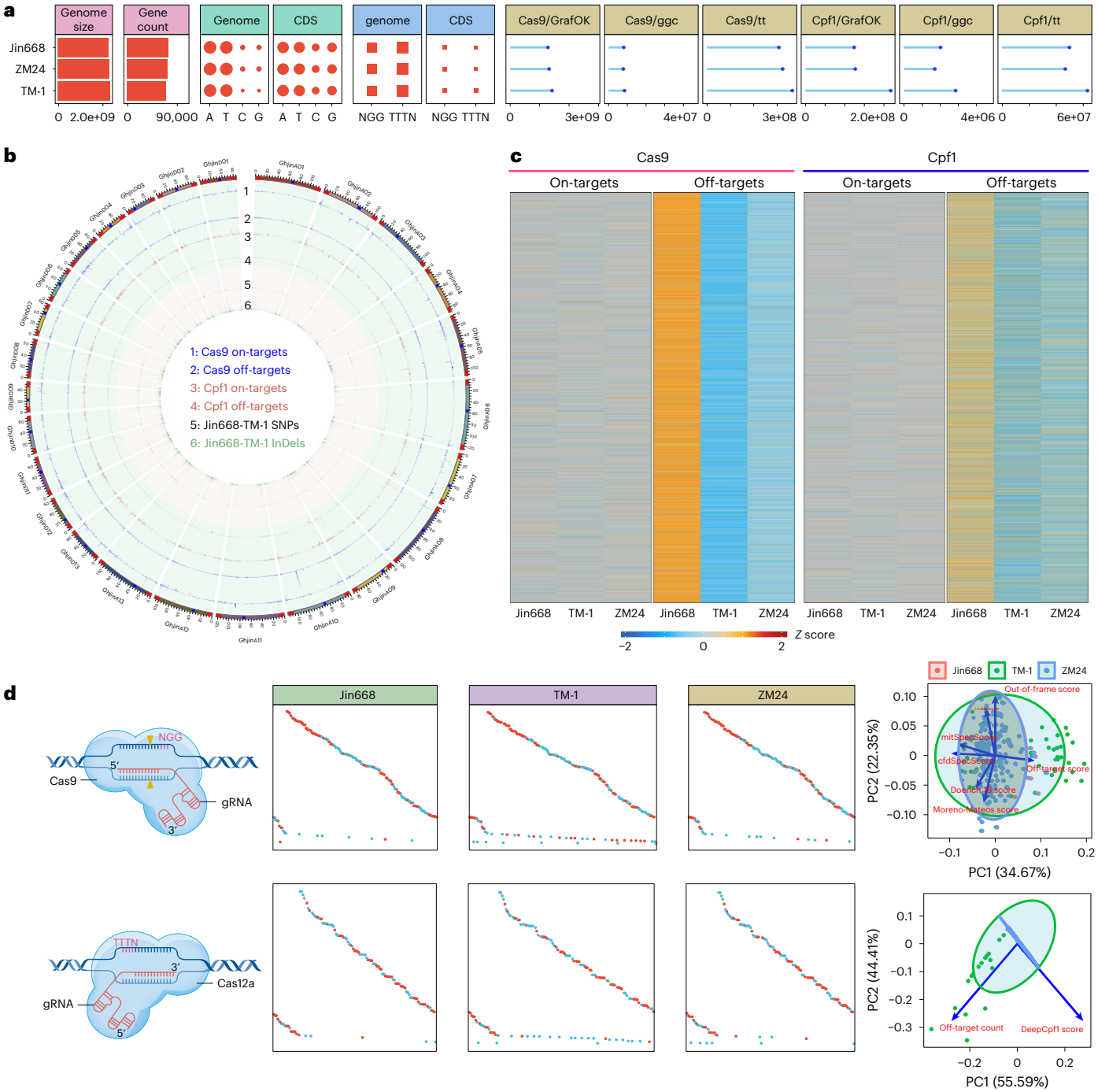


Fig. 7 | Cotton genetic variation substantially impacts the efficacy of CRISPR-based genome editing. a, The fraction of nucleotide bases within Jin668, ZM24, and TM-1, as well as the number of PAM sites, sgRNA targets and potent off-target sites for Cas9 and Cas12a enzymes. **b**, The circular diagram illustrates the positional correlation among the number of sgRNA targets, off-targets of 1:1 orthologous genes for Jin668 and TM-1, along with the existence of mutations between these two genotypes. The number of sgRNA targets and off-targets is obtained by subtracting the corresponding number in TM-1 from Jin668. **c**, Heatmap showing the number of differentially distributed sgRNA targets and off-targets of Cas9 and Cas12a across 1:1:1 orthologous genes for Jin668, ZM24 and TM-1. **d**, An example of a orthologous gene shows the impact of genomic variation on the accuracy of CRISPR-based genome editing.

included the absence of SE regulators in nonregenerable genotypes, defunctionalized or altered expression patterns in regenerative genotypes and potential regulatory roles of TEs through the insertion of promoters or coding[43]. In this study, TE polymorphisms associated with candidate SE-related genes suggest a potential regulatory role in cotton SE, warranting further investigation. Chromatin accessibility dynamics and differential expression of TFs and functional genes were prominent during SE in Jin668, contrasting with TM-1. These comprehensive comparisons of multi-level differences in the presence or absence of SE-restricted gene expression patterns provide specific directions and genetic resources for subsequent functional verification. Additionally, although this study identified the regeneration factor *AGL15* through single-gene knockout and overexpression, limitations remain, including low identification efficiency and persistent genotype-dependent constraints on regeneration. Therefore, future research employing artificial intelligence (AI)-powered multi-gene

manipulation could enable comprehensive analysis of the regeneration regulatory network.

Accurate sgRNA design is fundamental for CRISPR applications[39,44,45]. Our findings regarding the impact of genomic genetic variation on alterations of CRISPR specificity offer valuable insights for other species. Notably, CRISPR systems with smaller PAMs, like FnCas9-RHA (5′-YG-3′)[46], are more sensitive to genome selection and assembly quality. Therefore, constructing T2T-grade reference genomes for transformed genotypes is critical to ensure precise genome editing.

In conclusion, the availability of the Jin668 and YZ1 genomes enabled us to gain insight into the complete genome sequence of cotton. Moreover, our findings contribute to understanding the genetic regulation of SE, facilitating the construction of comprehensive regulatory networks. Thus, Jin668 genome serves as an essential resource for advancing functional genomics and genome editing applications in cotton, fostering both scientific progress and practical breeding advancements.

Online content

Any methods, additional references, Nature Portfolio 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/s41588-025-02258-3.

References

1. Bhatia, S., Sharma, K., Dahiya, R. & Bera, T. (eds). *Modern Applications of Plant Biotechnology in Pharmaceutical Sciences*, pp. 209–230 (Academic Press, 2015).
2. Zheng, Q. & Perry, S. E. Alterations in the transcriptome of Soybean in response to enhanced somatic embryogenesis promoted by orthologs of AGAMOUS-like15 and AGAMOUS-like18. *Plant Physiol.* **164**, 1365–1377 (2014).
3. Horstman, A., Bemer, M. & Boutilier, K. A transcriptional view on somatic embryogenesis. *Regeneration (Oxf.)* **4**, 201–216 (2017).
4. Wang, K. et al. The gene *TaWOX5* overcomes genotype dependency in wheat genetic transformation. *Nat. Plants* **8**, 110–117 (2022).
5. Chen, Z., Debernardi, J. M., Dubcovsky, J. & Gallavotti, A. Recent advances in crop transformation technologies. *Nat. Plants* **8**, 1343–1351 (2022).
6. Li, J. et al. Multi-omics analyses reveal epigenomics basis for cotton somatic embryogenesis through successive regeneration acclimation process. *Plant Biotechnol. J.* **17**, 435–450 (2019).
7. Iwase, A. et al. WIND1-based acquisition of regeneration competency in *Arabidopsis* and rapeseed. *J. Plant Res.* **128**, 389–397 (2015).
8. Lowe, K. et al. Morphogenic regulators *Baby boom* and *Wuschel* improve monocot transformation. *Plant Cell* **28**, 1998 (2016).
9. Debernardi, J. M. et al. A GRF–GIF chimeric protein improves the regeneration efficiency of transgenic plants. *Nat. Biotechnol.* **38**, 1274–1279 (2020).
10. Wang, M. et al. Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. *Nat. Genet.* **51**, 224–229 (2019).
11. Yang, Z. et al. Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. *Nat. Commun.* **10**, 2989 (2019).
12. Conover, J. L. & Wendel, J. F. Deleterious mutations accumulate faster in allopolyploid than diploid cotton (*Gossypium*) and unequally between subgenomes. *Mol. Biol. Evol.* **39**, msac024 (2022).
13. Sun, C. et al. Precise integration of large DNA sequences in plant genomes using PrimeRoot editors. *Nat. Biotechnol.* **42**, 316–327 (2024).
14. Chen, K., Wang, Y., Zhang, R., Zhang, H. & Gao, C. CRISPR/Cas genome editing and precision plant breeding in agriculture. *Annu. Rev. Plant Biol.* **70**, 667–697 (2019).
15. Wang, G. et al. Precise fine-turning of *GhTFL1* by base editing tools defines ideal cotton plant architecture. *Genome Biol.* **25**, 59 (2024).
16. Jiang, T., Zhang, X.-O., Weng, Z. & Xue, W. Deletion and replacement of long genomic sequences using prime editing. *Nat. Biotechnol.* **40**, 227–234 (2022).
17. Fernie, A. R. & Yan, J. *De novo* domestication: an alternative route toward new crops for the future. *Mol. Plant* **12**, 615–631 (2019).
18. Shoemaker, R., Couche, L. & Galbraith, D. Characterization of somatic embryogenesis and plant regeneration in cotton (*Gossypium hirsutum* L.). *Plant Cell Rep.* **5**, 178–181 (1986).
19. Jin, S. et al. Identification of a novel elite genotype for in vitro culture and genetic transformation of cotton. *Biol. Plant.* **50**, 519–524 (2006).
20. Wang, L. et al. The *GhmiR157a–GhSPL10* regulatory module controls initial cellular dedifferentiation and callus proliferation in cotton by modulating ethylene-mediated flavonoid biosynthesis. *J. Exp. Bot.* **69**, 1081–1093 (2017).
21. Xu, J. *GhL1L1* affects cell fate specification by regulating *GhPIN1*-mediated auxin distribution. *Plant Biotechnol. J.* **17**, 63–74 (2019).
22. Deng, J. et al. GhTCE1–GhTCEE1 dimers regulate transcriptional reprogramming during wound-induced callus formation in cotton. *Plant Cell* **34**, 4554–4568 (2022).
23. Yuan, J. et al. GhRCD1 regulates cotton somatic embryogenesis by modulating the GhMYC3–GhMYB44–GhLBD18 transcriptional cascade. *New Phytol.* **240**, 207–223 (2023).
24. Guo, H. et al. Somatic embryogenesis critical initiation stage-specific mCHH hypomethylation reveals epigenetic basis underlying embryogenic redifferentiation in cotton. *Plant Biotechnol. J.* **18**, 1648–1650 (2020).
25. Ge, X. et al. Efficient genotype-independent cotton genetic transformation and genome editing. *J. Integr. Plant Biol.* **65**, 907–917 (2023).
26. Liu, Y. et al. Cloning and preliminary verification of telomere-associated sequences in upland cotton. *Comp. Cytogenet.* **14**, 183–195 (2020).
27. Wang, P. & Wang, F. A proposed metric set for evaluation of genome assembly quality. *Trends Genet.* **39**, 175–186 (2023).
28. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
29. Vollger, M. R. et al. Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019).
30. Luo, S. et al. The cotton centromere contains a Ty3-Gypsy-like LTR retroelement. *PLoS ONE* **7**, e35261 (2012).
31. Gorinšek, B., Gubenšek, F. & Kordiš, D. A. Evolutionary genomics of chromoviruses in eukaryotes. *Mol. Biol. Evol.* **21**, 781–798 (2004).
32. Naish, M. et al. The genetic and epigenetic landscape of the *Arabidopsis* centromeres. *Science* **374**, eabi7489 (2021).
33. Schmitz, R. J., Grotewold, E. & Stam, M. *Cis*-regulatory sequences in plants: their importance, discovery, and future challenges. *Plant Cell* **34**, 718–741 (2021).
34. Zhu, X. et al. Single-cell resolution analysis reveals the preparation for reprogramming the fate of stem cell niche in cotton lateral meristem. *Genome Biol.* **24**, 194 (2023).
35. Braybrook, S. A. & Harada, J. J. *LECs* go crazy in embryo development. *Trends Plant Sci.* **13**, 624–630 (2008).
36. Ji, J. et al. *WOX4* promotes procambial development. *Plant Physiol.* **152**, 1346–1356 (2009).
37. Wang, F. et al. Chromatin accessibility dynamics and a hierarchical transcriptional regulatory network structure for plant somatic embryogenesis. *Dev. Cell* **54**, 742–757 (2020).

38. Izhaki, A. & Bowman, J. L. KANADI and Class III HD-Zip gene families regulate embryo patterning and modulate auxin flow during embryogenesis in *Arabidopsis*. *Plant Cell* **19**, 495–508 (2007).
39. Wang, G. et al. Development of an efficient and precise adenine base editor (ABE) with expanded target range in allotetraploid cotton (*Gossypium hirsutum*). *BMC Biol.* **20**, 45 (2022).
40. Li, C. et al. Targeted, random mutagenesis of plant genes with dual cytosine and adenine base editors. *Nat. Biotechnol.* **38**, 875–882 (2020).
41. Xue, C. et al. Tuning plant phenotypes by precise, graded downregulation of gene expression. *Nat. Biotechnol.* **41**, 1758–1764 (2023).
42. Xu, M., Du, Q., Tian, C., Wang, Y. & Jiao, Y. Stochastic gene expression drives mesophyll protoplast regeneration. *Sci. Adv.* **7**, eabg8466 (2021).
43. Zhang, L. et al. A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. *Nat. Commun.* **10**, 1494 (2019).
44. Qin, L. et al. High-efficient and precise base editing of C·G to T·A in the allotetraploid cotton (*Gossypium hirsutum*) genome using a modified CRISPR/Cas9 system. *Plant Biotechnol. J.* **18**, 45–56 (2020).
45. Jin, S. et al. Cytosine, but not adenine, base editors induce genome-wide off-target mutations in rice. *Science* **364**, 292–295 (2019).
46. Hirano, H. et al. Structure and engineering of *Francisella novicida* Cas9. *Cell* **164**, 950–961 (2016).
47. 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).
48. Chen, Z. J. et al. Genomic diversifications of five Gossypium allopolyploid species and their impact on cotton improvement. *Nat. Genet.* **52**, 525–533 (2020).
49. Sreedasyam, A. et al. Genome resources for three modern cotton lines guide future breeding efforts. *Nat. Plants* **10**, 1039–1051 (2024).
50. Han, J. et al. Rapid proliferation and nucleolar organizer targeting centromeric retrotransposons in cotton. *Plant J.* **88**, 992–1005 (2016).

Methods

Plant material

The highly homozygous *G. hirsutum* genotypes Jin668 and YZ1 were cultivated in the greenhouse in Huazhong Agricultural University for DNA extraction and sequencing. For Illumina, PacBio and ONT sequencing, young leaves were collected for high molecular weight DNA extraction, and then prepared and sequenced on the Illumina HiSeq 2500, PacBio Sequel II and Oxford Nanopore PromethION platform, respectively. For optical maps, the long, high-quality DNA extracted from young leaves according to the BioNano genomics protocol was labeled with the enzyme Nt.BssSI (New England Biolabs) and then loaded into the Saphyr chip for scanning. Hi-C library was constructed by chromatin extraction, digestion using DpnII (New England Biolabs), ligation, purification and fragmentation using standard protocols. Then, the libraries were sequenced with 150-bp paired-end reads using the Illumina HiSeq platform. Detailed library construction and sequencing methods are described in Supplementary Methods.

For the expression atlas sequencing of Jin668 and YZ1, diverse tissues, included seed, root, stem, leaf, anther, stigma, petal, sepal, bract, phloem, cotyledon and ovules at different developmental stages (0, 5, 10 and 20 days post anthesis), representing the major organ systems were collected and immediately frozen in liquid nitrogen, with two biological replications. Total RNA was extracted and purified, followed by DNase treatment. RNA-seq libraries were constructed and sequenced on the Illumina HiSeq 2500 platform with 150 bp paired-end sequences according to the manufacturer's protocol.

Processing of ONT and PacBio reads

For the ultralong ONT reads, fifteen cells of Nanopore ultralong reads were sequenced on the PromethION platform. Subsequently, the FAST5 files generated by Nanopore sequencers were converted to FASTQ format after base calling by using Guppy (v6.3.4)[51] with parameters '–c dna_r9.4.1_450bps_fast.cfg and mean_qscore_template 7'.

For genomic DNA associated with PacBio reads, the raw reads were imported into ccs (v6.0) to generate HiFi reads with the parameters '--min-passes 1 --min-rq 0.99 --min-length 100'.

Genome assembly and scaffolding

For Jin668 genome, the assembly was constructed using the NextDenovo (v2.5.0)[52]. First, the passed high-quality ONT reads were self-corrected using NextCorrect with parameters 'reads_cutoff:1k, seed_cutoff:80669' to obtain consistent sequences. Then, approximately 56 Gb (~24×) ultralong reads (>100 kb) were used to perform a de novo assembly via the NextGraph model, and the preliminary genome was assembled. To improve the accuracy of the assembly, the contigs were refined sequentially using ONT long reads (iterative correction three rounds), PacBio HiFi reads (iterative correction three rounds) and Illumina short reads (iterative correction four rounds) using NextPolish (v1.4.1)[53] with default parameters. The contigs were anchored into chromosomes using 3D-DNA (v180114) based on the Hi-C data.

For YZ1 genome, these long and highly accurate PacBio HiFi reads were assembled using Hifiasm (v0.18.1)[54] with default parameters to generate a draft contig genome. Subsequently, the Bionano optical maps were used to link the contigs into scaffolds using hybridScaffold.pl script in Bionano Solve Pipeline (v3.5.1_01142020) with parameter of '-B 2 -N 2'. Then, the hybrid scaffolds were further anchored into chromosomes using 3D-DNA (v180114) based on the Hi-C data and the Juicebox Assembly Tools (JBAT v1.8.8) were then applied for the manual correction of the order or orientation of several misassembled scaffolds on the basis of the Hi-C contact frequency.

Iterative pseudomolecule polishing and gap filling

As suggested in Col-CEN[32] and T2T-CHM13 (ref. 28) assembly, we corrected misassemblies and filled gaps in the Jin668 and YZ1 pseudomolecules with multiple iterations using ONT, HiFi and Illumina reads, sequentially. The detailed methodology is provided in the Supplementary Methods.

Identification, patching and polishing of telomeres

Using an in-house Python script, the telomeric repeat sequences in the assembled genome were identified based on previously reported telomeric repeats (AAACCCT/AGGGTTT)[26] at first.

To recall missing telomeres in certain chromosome arms, we performed multiple strategies to remap all available primary data, including ONT and HiFi, to the assembled genome. First, ONT reads containing telomeric sequences were identified using the script as previously mentioned. Then, these associated ONT reads were run to generate a consensus sequence using Medaka (v1.6.0) and manually patched into the assembly. Additionally, HiFi reads with confirmed telomeric repeats were identified using the same script and further polished with Racon (v1.6.0). Additionally, Teloclip (v0.1.0) was also employed to extract telomere sequences that contain species-specific repeating units from ONT and HiFi reads, as well as assembled contigs using Hifiasm (v0.18.1)[54] with combined ONT and HiFi reads. These sequences were subsequently mapped to the assembled genome to resolve missing telomeres.

Identification and copy number estimation of rDNA arrays

The rDNA arrays were preliminarily identified by BLAST with reported 5S rRNA monomer (GenBank ID: U32085), 18SrRNA (GenBank ID: U42827), ITS1_5.8SrRNA_ITS2 sequences (GenBank ID: KC404827) and 28SrRNA (GenBank ID: XR_005924313.1) repeats against the Jin668 genome. Specifically, for 5S rDNA (usually around 300 bp), the pyTanFinder[55] with parameters '--minMonLength 100 --maxMonLength 500' was also set to identify and cluster the repetitive elements of different monomer sizes in the Jin668 genome. Then, the repetitive units within the 5S and 45S rDNA regions were aligned with the nucleotide sequence database of National Center for Biotechnology Information (NCBI) to determine whether they are annotated as rDNA. In addition, the copy number of 5S and 45S rDNAs in the Jin668 genome was also estimated by the BLAST-based method, using both ONT ultralong and PacBio HiFi data to verify the accuracy of assembly. Detailed methods are described in the Supplementary Methods.

Annotation and validation of centromeres

Active centromere locations were determined by identifying the CENH3 ChIP–seq-enriched regions in the final assembly using input DNA as a control. A peptide was synthesized to the 12 most C-terminal amino acids (MSRTKHTAAKKP) of the Jin668 and YZ1 CenH3 protein conjugated with a cysteine. The peptide was then conjugated to keyhole limpet hemocyanin and injected into two rabbits to raise the polyclonal antibody. The antisera were affinity purified to obtain specific antibodies, which were subsequently tested for binding specificity (Supplementary Methods and Supplementary Fig. 5b) and then used for ChIP experiments following a published protocol[50]. Briefly, the harvested leaf tissue was fixed with 1% formaldehyde and then ground into a fine powder. The nuclei were isolated using NIB buffer (10 mM Tris–HCl pH 9.5, 80 mM KCl, 10 mM EDTA pH 8.0, 0.5 M sucrose, 1 mM spermidine, 200 µl 1% proteinase inhibitor). Isolated nuclei were sonicated into 100–300 bp fragments. After sonication, the chromatin solution was purified by centrifugation and equally divided into two tubes. CENH3 antibody at a concentration of 2 mg ml^{-1} was added to one tube for ChIP, while the other tube without the antibody served as the input control. The DNAs from both the ChIP and input were then used for library construction according to the protocol provided by Illumina, and sequenced using the Illumina HiSeq 2500 platform.

Both the ChIP reads and input samples reads were trimmed with Trim Glore (v0.6.7) and aligned to assembled Jin668 or YZ1 genome with BWA-MEM (v0.7.17) with default parameters. PCR duplicate-like

redundancies were removed using Picard (v2.23.9) with default parameters. Epic2 (ref. 56) was used to call peaks, with CENH3 ChIP–seq alignment as the treatment and input sample alignment as the control. Parameters included a MAPQ threshold of 20, an effective genome fraction, a bin size of 5,000 and a gap size of 0. Especially, the effective genome fraction of the assembly genome was calculated as the fraction of unique 150-mers over total 150-mers using Jellyfish (v2.3.0)[57] (-m 150 -out-counter-len 1 -counter-len 1). Based on the Gapless assembly of the maize genome[58], the coordinates of active centromeres were identified as islands with a score above 250 and a fold change higher than 4.

The coordinates of centromere positions for each chromosome were calculated by normalizing RPKM (reads per kilobase of transcript per million fragments sequenced) values from the ChIP data against the genomic input in 5 kb windows with deepTools (v3.5.0)[59]. Default options were used except for the following parameters: --ignoreDuplicates --scaleFactorsMethod None --normalizeUsing RPKM. Finally, the coordinates of the centromere were identified with a ratio of enriched islands above 2.5 and merged with a distance interval of 1 Mb using bedtools (v2.27)[60]. Subsequently, the centromere locations were manually adjusted and confirmed using Integrative Genomics Viewer.

To verify the quality of centromere regions of Jin668, HiFi data was aligned to the centromeric sequences using pbmm2 (v1.16.0) with parameters '--log-level DEBUG -- preset SUBREAD --min-length 5000' and the frequency of secondary alleles was generated and plotted using NucFreq (v0.1). Additionally, TandemTools[61] was also used to validate the centromeric assembly using ONT reads and the coverage of base-calling quality across the whole centromere was plotted. After quality verification, the centromeric tandem array on each chromosome was identified using the pyTanFinder[55]. The most frequent consensus sequence of the tandem repeats that were present in all chromosomes was considered the centromeric repeat unit.

To resolve the sequence of the centromeric transposable elements, we used the Extensive de novo TE Annotator (EDTA; v2.0.0)[62] and DANTE-Protein Domain Finder (v0.2.5)[63] to annotate all centromeric sequences. Detailed methods are described in the Supplementary Methods. The Structure of centromeric was shown using heatmap with pairwise sequence identity between all nonoverlapping 5-kb regions through PT2T_StainedGlass modified from mrvollger/StainedGlass[64].

FISH assay

To verify the specific repetitive sequence of Jin668, FISH assay was conducted using conserved sequences located within the 5S and 45S rDNA arrays, centromere-specific repeat array, and telomere regions as probes, following the previously published protocol[65]. Detailed methods are described in Supplementary Methods.

Genome mappability

In an effort to assess the mappability of Jin668 and YZ1 assembly, we used GenMap (v1.3.0) to find positions of unique *k*-mers (set to 50-, 100-, 150-, 200- and 300-mers) in each genome position, with up to *e* mismatches (zero mismatches were permitted). The (*k*,*e*)-mappability for a given position represents the reciprocal value of the frequency with which the *k*-mer occurs in the genome. Chromosome-scale profiles were generated by calculating mean (*k*,*e*)-mappability values within adjacent 10-kb genomic windows. Genomic mappability with different *k*-mers and CENH3 $\log_2$(ChIP/input) peaks were visualized through R packages karyoploteR[66].

Assembly validation and genome quality evaluation

We validated the final assembly using a mixture of techniques and sequencing data, including detection of organelle genome, statistics of genome GC content, alignment-based validation, BioNano optical mapping validation, Hi-C validation, *k*-mer based validation and assessment of assembly completeness, as described in detail in the Supplementary Methods.

Annotation of repetitive sequences and gene models

The annotation of repetitive DNA followed both homology-based prediction and de novo identification of repeats as previously described[67]. In brief, a de novo repeat library was constructed using RepeatModeler (v2.0.4)[68] and EDTA (v2.0.0)[62]. Then, we adapted RepeatMasker (v4.0.7)[68] to perform a homology-based repeat search and masked throughout the Jin668 and YZ1 genomes using both the Repbase (Repbase21.08)[69] and the nonredundant de novo repeat library.

A well-developed combination of strategies integrating homology-based prediction, RNA-sequencing-assisted prediction and ab initio prediction was used for gene model prediction using genome in which all repetitive regions had been soft-masked. The methods of gene annotation are described in detail in the Supplementary Methods.

Genome comparison and structural variation analysis

The orthologous gene identification, synteny analysis and visualization were performed using JCVI (https://github.com/tanghaibao/jcvi) with default parameters. Only the two-way best gene pairs with $P < 1 \times 10^{-20}$ were retained.

For comparative genome analysis, genome sequences of Jin668, YZ1, ZM24 and two Upland cotton ancestors A_2 (*G. arboreum*)[47] and D_5 (*G. raimondii*)[70] were aligned to the TM-1 reference genome[49] using minimap2 (v2.16-r922)[71] with the parameter settings '-x asm5 -eqx'. Then, SyRI (v1.6)[72] was used to identify genome-based synteny and structural variations for further analysis. Functional effects of SNPs and InDels under selection were predicted using SnpEff (v4.3)[73]. PAVs were extracted by scanPAV[74] with default parameters, and the resulting PAVs shorter than 1,000 bp were filtered out as noise.

The genome resequencing data of each accession were mapped to the TM-1, Jin668, YZ1 and ZM24 reference genomes using BWA (v0.7.17)[75] with default parameters. Then, the reads with the mapping quality value < 20 were removed by Samtools (v1.6)[76]. The Picard program (v2.23.9) was used to mark duplicative reads, and Genome Analysis Toolkit (GATK; v4.1)[77] was employed to call SNPs and InDels. The high-confidence SNVs that filtered with parameters 'QUAL < 30.0 || MQ < 50.0 || QD < 2' were annotated using SnpEff (v4.3)[73].

To identify the real orthologous gene pairs and exclude potential annotation discrepancies and assembly quality issues, we performed a homologous alignment of the protein sequences from the reference genome to the query genome using TBLASTN (-evalue 1e-5 -max_target_seqs 20). Subsequently, GeneWise (v2.4.1)[78] with default parameters was used to predict the exact gene structure of the corresponding genomic region on each conjoined hit.

Gene expression and co-expression analysis of RNA-seq data

The RNA-sequencing reads were removed of adapters and trimmed for low-quality bases using Trimmomatic (v0.39)[79]. The clean reads were then mapped to the reference genome using HISAT2 (v2.2.1)[80] with default parameters. The expression level (transcripts per million (TPM)) of genes was calculated by StringTie (v2.1.4)[81]. A gene was considered to be expressed if its TPM > 0. Subsequently, differentially expressed genes were identified by using the DESeq2 package[82] with at least a twofold change in expression and a false detection rate (FDR) value of less than 0.05.

To identify relationships between expressed genes, all RNA-seq data from the samples described above and required genes with a TPM value of ≥ 1 in at least one sample were used to perform a weighted gene co-expression analysis using the R package WGCNA[83]. Cytoscape (v3.10.3)[84] was used to visualize the network.

The GO and KEGG enrichment analysis of differentially expressed genes or the co-expression gene sets were performed using KOBAS (v3.0)[85].

Library construction of cotton hypocotyl for ATAC–seq

For direct SE, sterilized seeds of Jin668, YZ1, TM-1 and ZB1092 (an Upland cotton genotype incapable of regeneration from somatic cells; Supplementary Fig. 21) were germinated on MS medium and grown in the dark for 7 days. After that, use a sterilized scalpel to cut the hypocotyl into 1 cm segments and place them on a solid auxin 2,4-dichlorophenoxyacetic acid (2,4-D) MSB medium, and start timing (named 0 HAI). For RNA-seq and ATAC–seq, two biological replicates of approximately 20 hypocotyl segments from TM-1, ZB1092, YZ1 and Jin668 were collected at 0, 0.5, 1, 6 and 12 HAI (Supplementary Note 7). The methods for library construction in ATAC–seq are described in detail in the Supplementary Methods.

ATAC–seq data analyses

ATAC–seq data analysis was performed following published methods with some modifications[86]. Briefly, reads that were trimmed and checked using fastp (v0.23.1)[87] were mapped to the reference genomes (Jin668, YZ1 and TM-1) using Bowtie2 (v2.5.2) with parameters '--local -k 20 --very-sensitive -X 2000 --no-mixed --no-discordant'. The mapped reads from mitochondrial and chloroplast DNA were manually excluded. Samblaster (v0.1.26)[88] with the removeDups parameter was used to remove PCR duplicates and keep mapped reads of high mapping quality. Peak calling was performed using MACS2 (v2.2.7.1)[89] with parameters '--nomodel --keep-dup all -B --SPMR --call-summits --shift -100 --extsize 200 --nolambda'.

To ensure robust data quality, ENCODE-recommended metrics were calculated. The signal portion of tags score, reflecting the percentage of reads aligned to identified accessible regions, was calculated by subsampling twenty million reads. The fraction of reads in peaks and correlation of biological replicates were calculated via Irreproducible Discovery Rate (IDR)[90]. Additionally, 'PlotProfile' and 'PlotHeatmap' functions from deeptools were used for assessment of sample quality and associated genomic regions. The 'multiBamSummary' function in deeptools[59] was used to compute read coverages for genomic regions, followed by principal component analysis plot using DESeq2 (ref. 82) in R.

The peaks called by MACS2 (v2.2.7.1) and differential peaks ($|\log_2(\text{fold change})| \geq 1$ and false discovery rate < 0.05) identified by DiffBind (v2.14.0) were annotated through 'annotatePeaks.pl' script in HOMER (v4.11.1)[91]. Likewise, the differential peaks regions were scanned for motif enrichment using monaLisa[92], which calls 'findMotifsGenome.pl' function provided by HOMER (v4.11.1)[91] with a custom nonredundant motif dataset as described below. It is worth noting that due to the absence of some SE-related TF motifs such as *WOX*, *BBM*, *LEC1*, *ERF59*, *ERF107* and *PLT5* in either the JASPAR2022 (ref. 93) or HOMER database, we collected all available SE-related motifs from Plant Transcription Factor Database (PlantTFDB; v5.0) and converted them to Homer format using TFBStools[94]. Then merged and removed the redundancy with the motifs provided by the JASPAR202292 and HOMER databases. Moreover, 'plotMotifHeatmaps' function in monaLisa[92] was also used to plot the motif enrichments in the form of heatmap.

Target gene knockout and overexpression bioassays

The sgRNAs of all candidate genes were designed using CRISPOR (v3.1)[95] program with the Jin668 genome as reference. Each sgRNA was cloned into CRISPR/Cas9 vector pRGEB32-GhU6.7 and introduced into *Agrobacterium* strain GV3101 by electroporation, and then carried out genetic transformation via *Agrobacterium*-mediated with Jin668 and TM-1 as transformation receptor according to our previous publications[15]. The full-length coding sequence of candidate genes was amplified from cDNA of Jin668 and cloned into the overexpress vector PK2GW7, then performed *Agrobacterium*-mediated genetic transformation as described in CRISPR/Cas9 experiment.

The CPR was calculated as the fold change in weight gained of explants at 20 and 60 days post-induction. Three biological replicates were included, and each replicate represented at least five culture dishes with more than 20 explants per dish.

Genome-wide sgRNA on-targeting and off-targeting identification

Target variation analysis was performed using CRISPOR (v3.1)[95], a tool using input sequences to find guide RNAs and provide multiple evaluation scores for on-target and off-target, with default parameters. Briefly, genome sequences from selected species were used to build a database, and the longest isoform was extracted and split into an exome for input into CRISPOR (v3.1). The specificity (mitSpecScore and cfdSpecScore) and predicted efficiency (Doench '16 and Moreno-Mateos score for Cas9 sgRNA and DeepCpf1 for Cpf1 sgRNA) scores were used to evaluate the accuracy and editing efficiency of the on-target. While the outcome (out-of-frame score and Lindel score) and off-target mismatch counts with a maximum of four mismatches were used to evaluate the potential off-targeting sites. Additionally, in the CRISPOR output, high-quality sgRNAs were labeled as 'GrafOK', while inferior sgRNAs, characterized by TT- and GCC-motif endings, were labeled as 'tt' and 'ggc', respectively.

Statistical analyses

Details of the statistics of Figs. 3f and 6b and Extended Data Fig. 8b are provided in the figure legends. All statistics were performed using two-sided Student's *t*-test in R, with $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, unless otherwise indicated.

Reporting summary

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

Data availability

The T2T-Jin668 and YZ1 genome assemblies and annotation data are available at NCBI (PRJNA874817 and PRJNA960814) and T2TCotton-Hub (http://jinlab.hzau.edu.cn/T2TCottonHub/ or http://cotton.hzau.edu.cn/T2TCottonHub/). The raw sequencing data used for de novo whole-genome assembly of Jin668 and YZ1 are available in NCBI under BioProjects PRJNA874817 and PRJNA960814, respectively. RNA-seq of Jin668 and YZ1 are available in NCBI under BioProjects PRJNA874819 and PRJNA960820, respectively. ATAC–seq data of Jin668 is available in NCBI under BioProjects PRJNA960832. ATAC–seq and RNA-seq data for TM-1 during SE are available in NCBI under BioProjects PRJNA960828 and PRJNA960825, respectively. ATAC–seq and RNA-seq data for YZ1 during SE are available in NCBI under BioProjects PRJNA1059614 and PRJNA1059613, respectively. ATAC–seq and RNA-seq data for ZB1092 during SE are available in NCBI under BioProjects PRJNA1059611 and PRJNA1059609, respectively. The ChIP–seq data for Jin668 and YZ1 are uploaded to BioProjects PRJNA1079680 and PRJNA1079682, respectively. Seeds of Jin668 and YZ1 used in this study can be obtained from the corresponding author upon request. The reference genome assembly and annotation files of TM-1 (v3.1) used in this study were downloaded from https://phytozome-next.jgi.doe.gov/ and are also accessible from T2TCottonHub. Additionally, all available SE-related motifs were obtained from Plant Transcription Factor Database (v5.0; https://planttfdb.gao-lab.org/). Further details on data accessibility are outlined in the Supplementary Methods and Methods. Source data are provided with this paper.

Code availability

All original codes used in this article are available via Zenodo at https://doi.org/10.5281/zenodo.15035095 (ref. 96) and GitHub (https://github.com/tiramisutes/T2T-Cotton-Genomes).

References

51. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford nanopore sequencing. *Genome Biol.* **20**, 129 (2019).
52. Hu, J. et al. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biol.* **25**, 107 (2024).
53. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2019).
54. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
55. Kirov, I., Gilyok, M., Knyazev, A. & Fesenko, I. Pilot satellitome analysis of the model plant, *Physcomitrella patens*, revealed a transcribed and high-copy IGS related tandem repeat. *Comp. Cytogenet.* **12**, 493–513 (2018).
56. Stovner, E. B. & Sætrom, P. epic2 efficiently finds diffuse domains in ChIP–seq data. *Bioinformatics* **35**, 4392–4393 (2019).
57. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764 (2011).
58. Liu, J. et al. Gapless assembly of maize chromosomes using long-read technologies. *Genome Biol.* **21**, 121 (2020).
59. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
60. Quinlan, A. R. BEDTools: the Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* **47**, 11.12. 1–11.12. 34 (2014).
61. Mikheenko, A., Bzikadze, A. V., Gurevich, A., Miga, K. H. & Pevzner, P. A. TandemTools: mapping long reads and assessing/improving assembly quality in extra-long tandem repeats. *Bioinformatics* **36**, i75–i83 (2020).
62. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275–292 (2019).
63. Novák, P., Neumann, P., Pech, J., Steinhaisl, J. & Macas, J. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29**, 792–793 (2013).
64. Vollger, M. R., Kerpedjiev, P., Phillippy, A. M. & Eichler, E. E. StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps. *Bioinformatics* **38**, 2049–2051 (2022).
65. Peng, R. et al. Evolutionary divergence of duplicated genomes in newly described allotetraploid cottons. *Proc. Natl Acad. Sci. USA* **119**, e2208496119 (2022).
66. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).
67. Hu, L. et al. The chromosome-scale reference genome of black pepper provides insight into piperine biosynthesis. *Nat. Commun.* **10**, 4702 (2019).
68. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457 (2020).
69. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
70. Paterson, A. H. et al. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* **492**, 423–427 (2012).
71. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
72. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
73. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; *iso-2*; *iso-3*. *Fly (Austin)* **6**, 80–92 (2012).
74. Giordano, F., Stammnitz, M. R., Murchison, E. P. & Ning, Z. scanPAV: a pipeline for extracting presence–absence variations in genome pairs. *Bioinformatics* **34**, 3022–3024 (2018).
75. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
76. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
77. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
78. Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. *Genome Res.* **14**, 988–995 (2004).
79. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
80. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
81. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
82. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
83. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
84. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
85. Bu, D. et al. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res.* **49**, W317–W325 (2021).
86. Grandi, F. C., Modi, H., Kampman, L. & Corces, M. R. Chromatin accessibility profiling by ATAC–seq. *Nat. Protoc.* **17**, 1518–1552 (2022).
87. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
88. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).
89. Zhang, Y. et al. Model-based analysis of ChIP–seq (MACS). *Genome Biol.* **9**, R137 (2008).
90. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. Preprint at https://arxiv.org/abs/1110.4705 (2011).
91. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
92. Machlab, D. et al. monaLisa: an R/Bioconductor package for identifying regulatory motifs. *Bioinformatics* **38**, 2624–2625 (2022).
93. Castro-Mondragon, J. A. et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **50**, D165–D173 (2021).
94. Tan, G. & Lenhard, B. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics* **32**, 1555–1556 (2016).

95. Concordet, J.-P. & Haeussler, M. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res.* **46**, W242–W245 (2018).
96. Xu, Z. Scripts used in 'Genome assembly of two allotetraploid cotton germplasms reveals mechanisms of somatic embryogenesis and enables precise genome editing'. *Zenodo* https://doi.org/10.5281/zenodo.15035095 (2025).

Acknowledgements

The study was supported by grants from the National Science Fund for Distinguished Young Scholars (32325039 to S.J.) and the Young Scientists Fund under the National Natural Science Foundation of China (32201856 to Z.X.). The computations in this paper were run on the bioinformatics computing platform of the National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University.

Author contributions

S.J. and X. Zhang designed and supervised the research. Z.X. and G.W. collected materials for genome and transcriptome sequencing. Z.X. and G.W. collected materials for ATAC–seq. Z.X. and R.W. performed bioinformatics analysis. Y.L. and R.P. performed the FISH experiment. G.W. and X.Z. performed the gene editing experiments. M.W., L.T. and L.Z. contributed to the project discussion. Z.X. and G.W. wrote the manuscript with input from all other authors. S.J., K.L., X. Zhang and M.W. edited the manuscript. All authors have read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

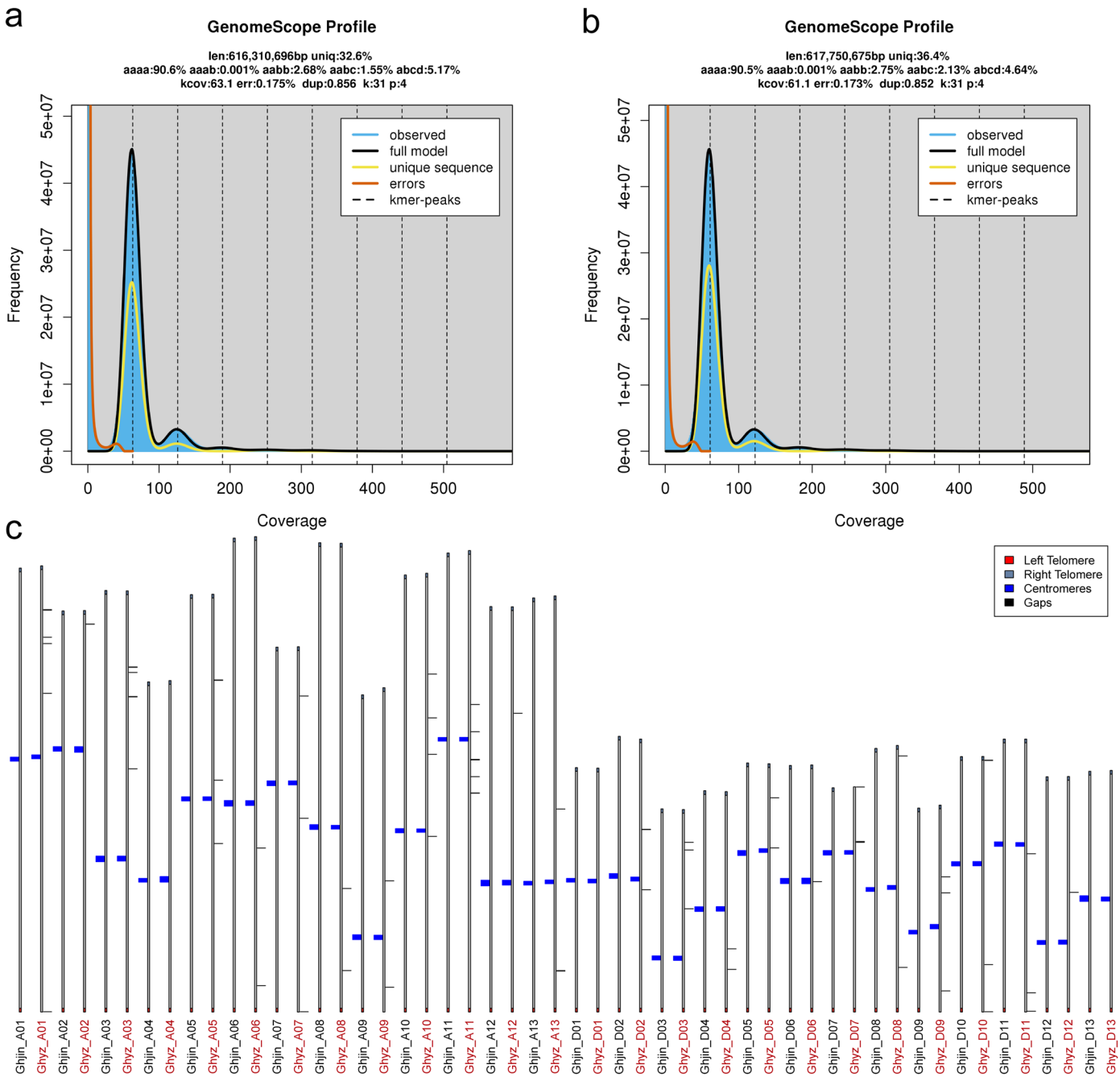
Extended data is available for this paper at https://doi.org/10.1038/s41588-025-02258-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-025-02258-3.

Correspondence and requests for materials should be addressed to Maojun Wang, Xianlong Zhang or Shuangxia Jin.

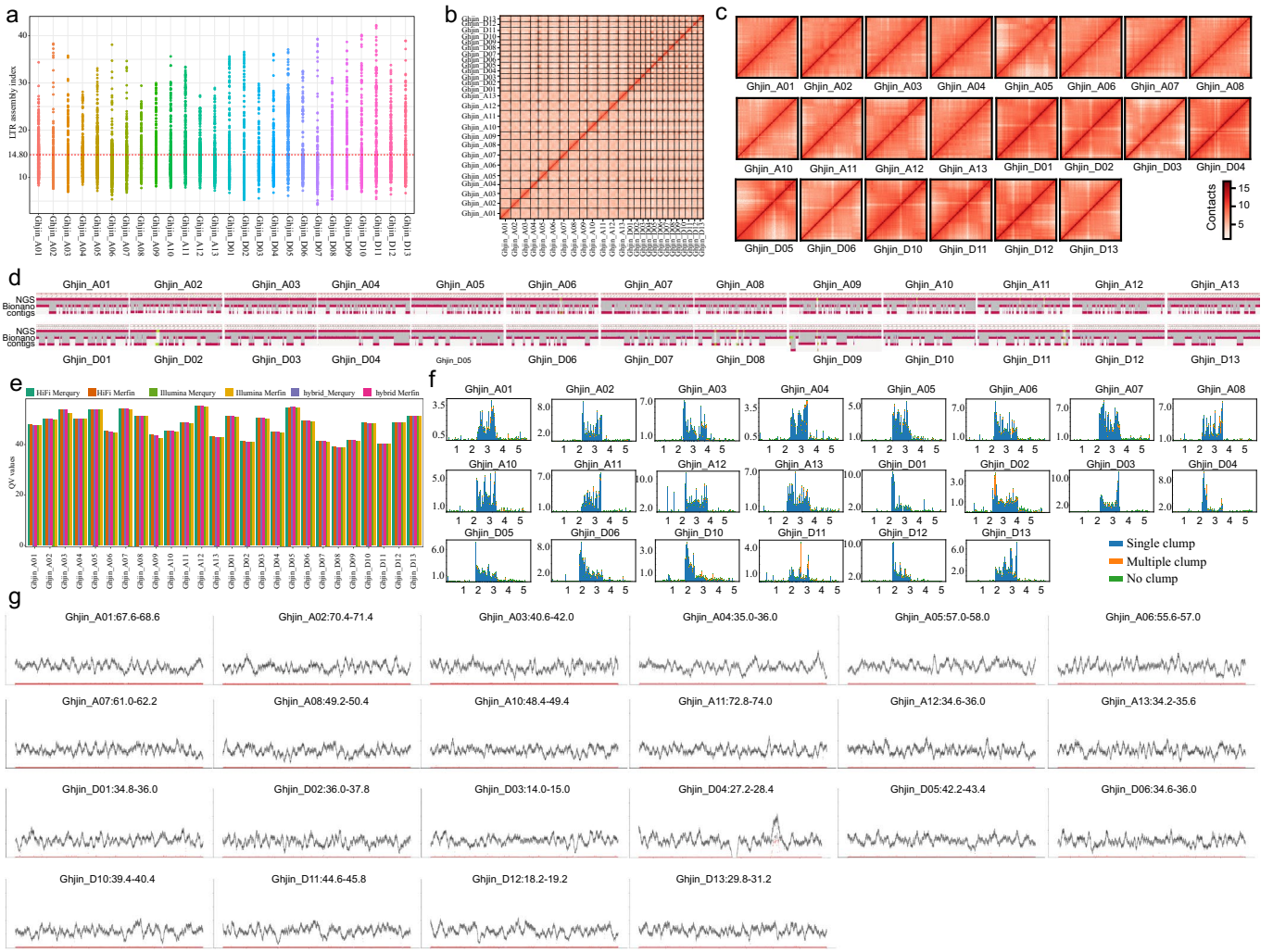
Peer review information *Nature Genetics* thanks Amanda Hulse-Kemp, Jinsheng Lai and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41588-025-02258-3



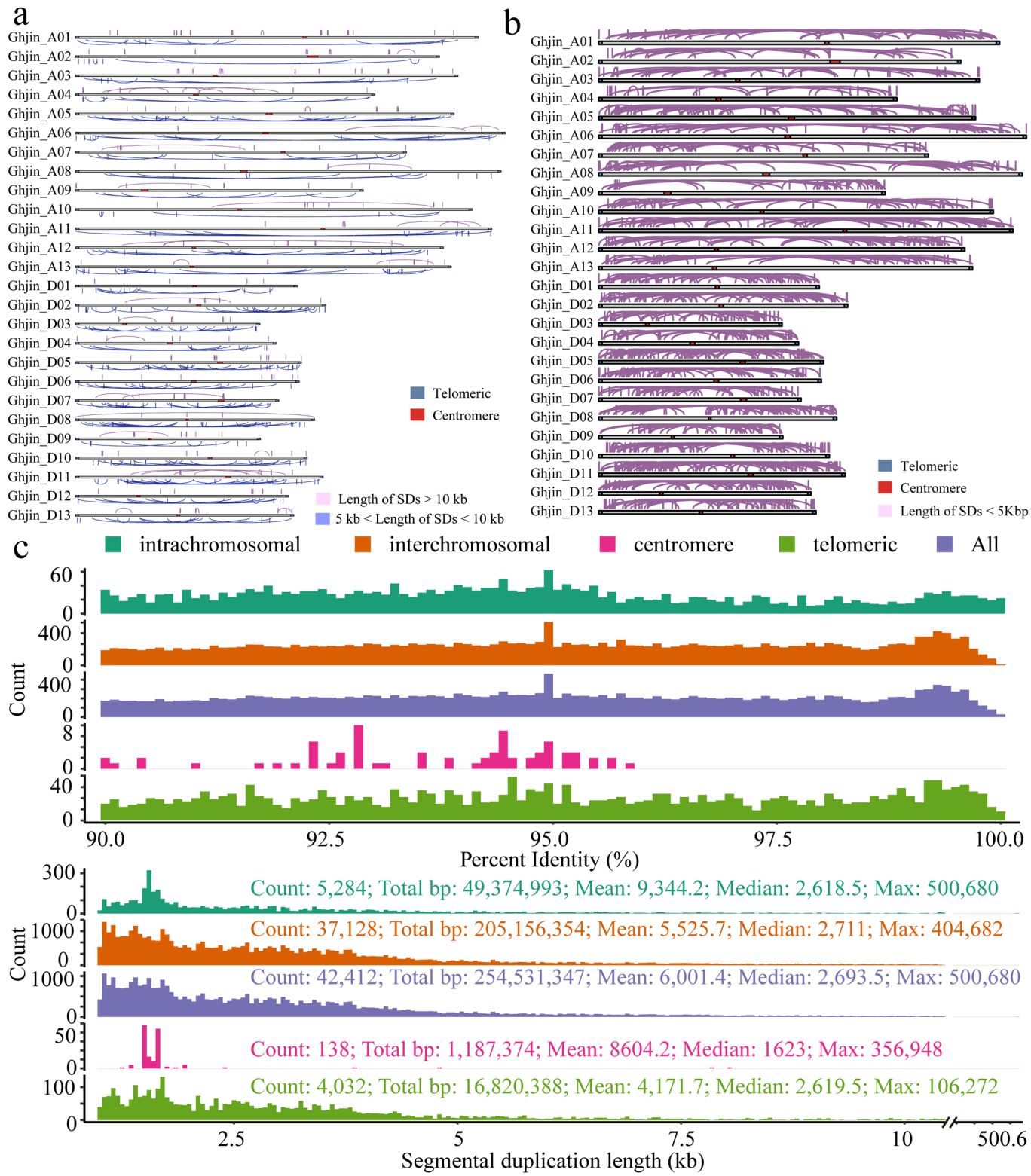
Extended Data Fig. 1 | Genomic characterization of Jin668 and YZ1. The k-mer (31-mer) frequency distribution was estimated by GenomeScope2.0 analysis of Jin668 (**a**) and YZ1 (**b**) genome. **c**, The distribution of telomeres, centromeres, and gaps regions in the genomes of Jin668 and YZ1.

https://doi.org/10.1038/s41588-025-02258-3



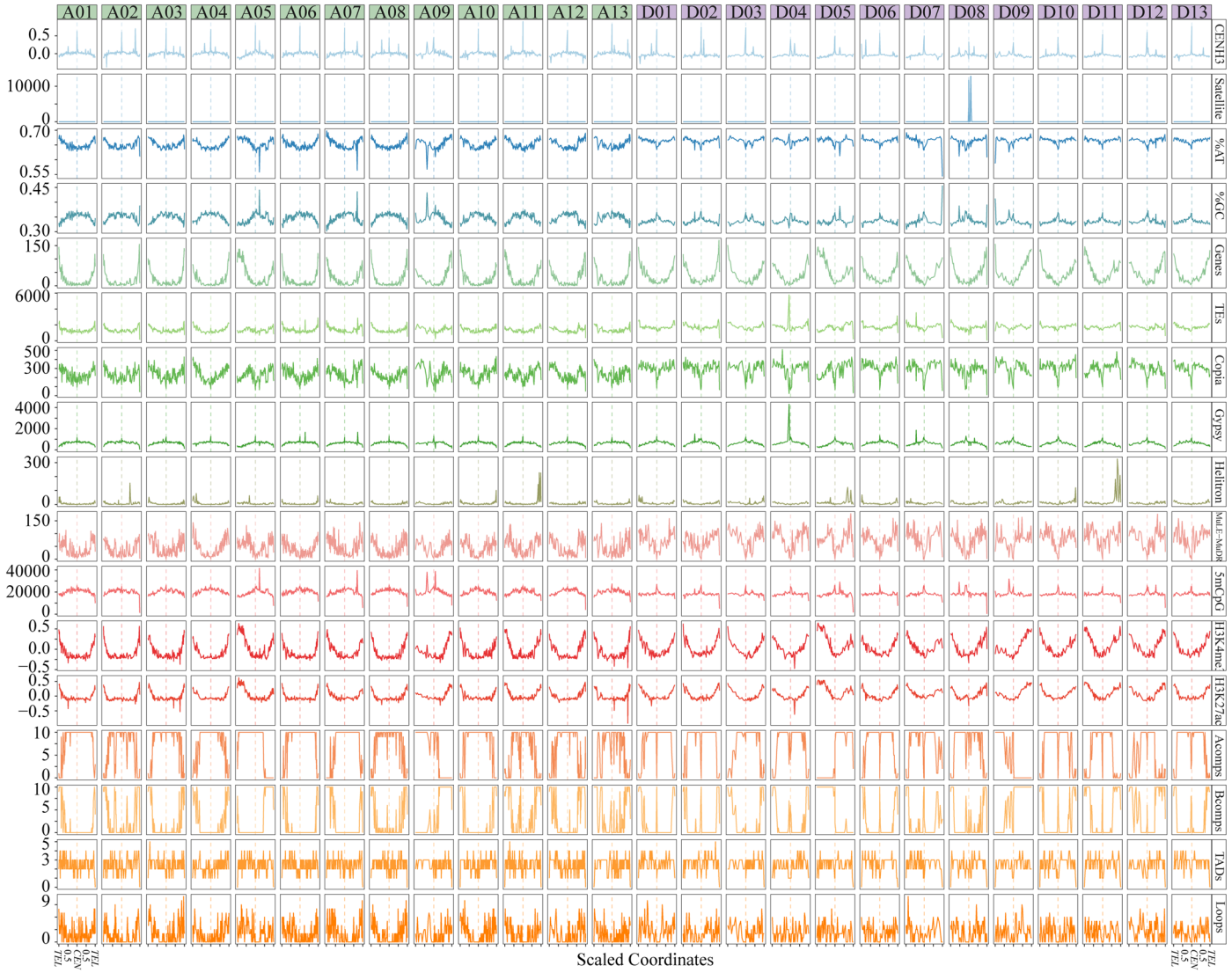
Extended Data Fig. 2 | Completeness and accuracy validation of chromosome arms and centromeres in the Jin668 genome. a, Assessment of the Jin668 genome assembly using LTR Assembly Index (LAI). The x-axis lists chromosomes and y-axis represents LAI values. The red dashed line represents the mean value. **b**,**c**, Hi-C map of the Jin668 genome showing genome- and chromosome-wide all-by-all interactions. **d**, Bionano *de novo* assembly contigs were mapped to the Jin668 reference assembly. **e**, The distribution of QV values calculated from Merqury and Merfin on different chromosomes. **f**, The distribution of different types of unique k-mers along the centromeric regions in chromosomes of Jin668. Each bar shows the number of different types of k-mers in a bin of length 20 Kb. The blue bars represent single-clump k-mers, which suggest a good base-level quality. While the orange (multiple-clumps) and green (no-clumps) bars suggest a low base-level quality in the region. The x-axis unit is megabases (Mb), and the y-axis represents values multiplied by 10^3. **g**, NucFreq plot of centromeric regions in chromosomes of Jin668. HiFi coverage depth (black) along with secondary allele frequency (red) for all centromeres and surrounding regions. The x-axis represents chromosome positions in megabases (Mb), and the y-axis shows HiFi depth, ranging from 0 to 100.

https://doi.org/10.1038/s41588-025-02258-3



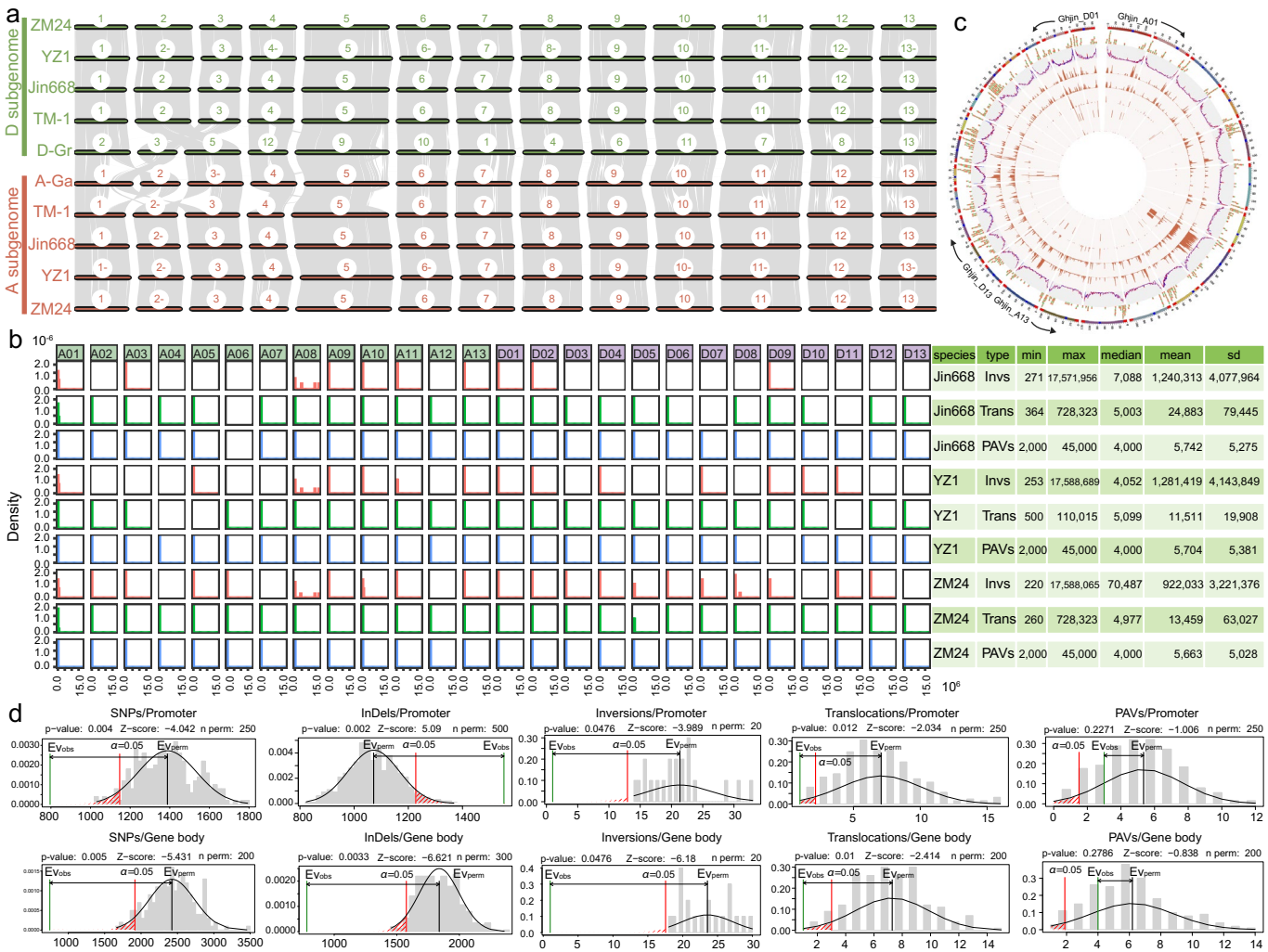
Extended Data Fig. 3 | Segmental duplication (SD) content of the Jin668 genome. a,**b**, The intrachromosome SDs with a length greater or less than 5 kb. **c**, Comparison of SD length and identity in different regions of the genome. The identity (top) and length (bottom) of SDs across commonly delineated regions of the genome (colors).

https://doi.org/10.1038/s41588-025-02258-3



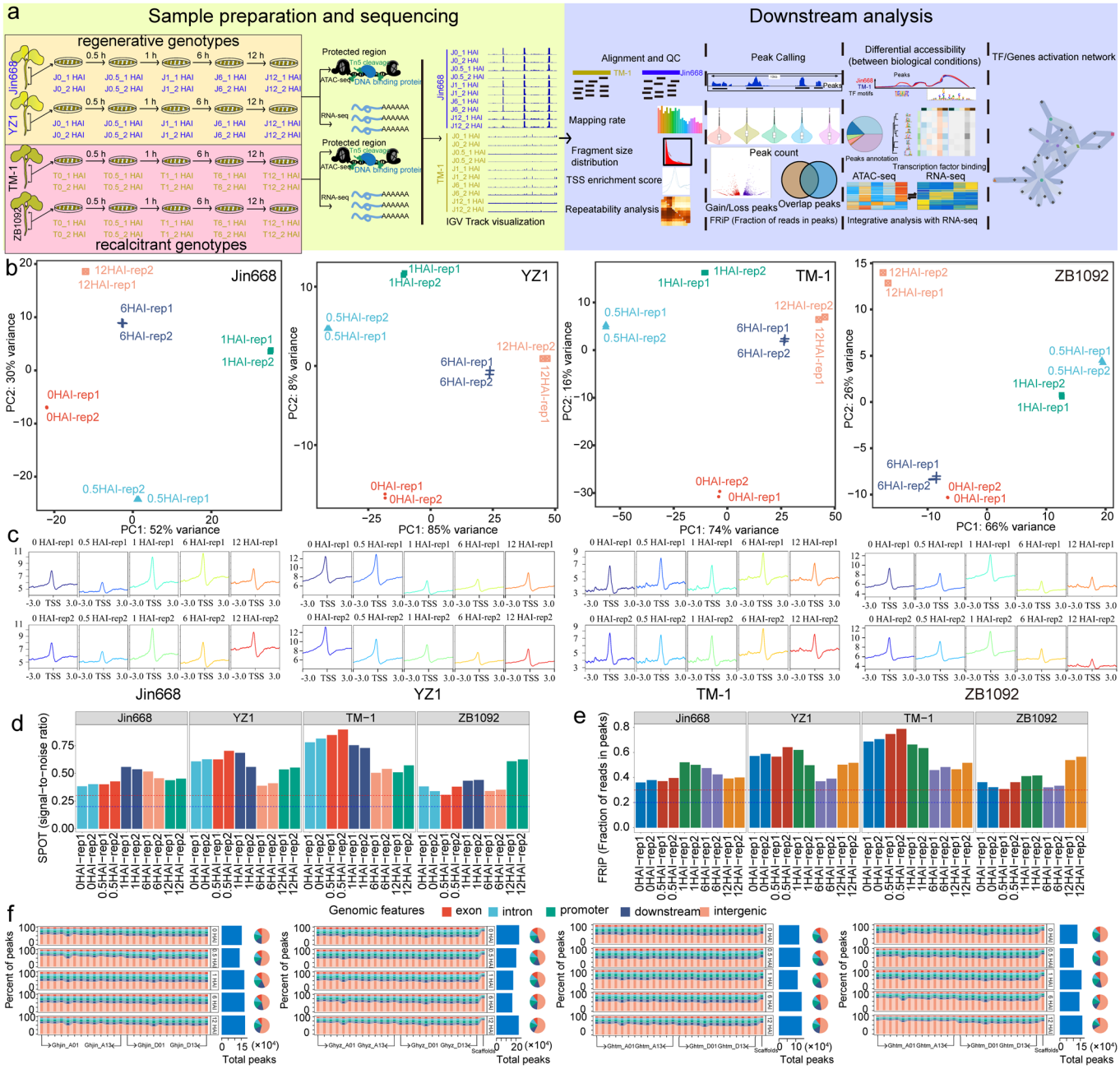
Extended Data Fig. 4 | The percentage of AT/GC base composition, TEs, 5mC DNA methylation, histone modification and 3D genome architecture within the centromeres. Quantification of genomic features plotted along chromosome arms that were proportionally scaled between telomeres (TEL) and centromere midpoints (CEN).

https://doi.org/10.1038/s41588-025-02258-3



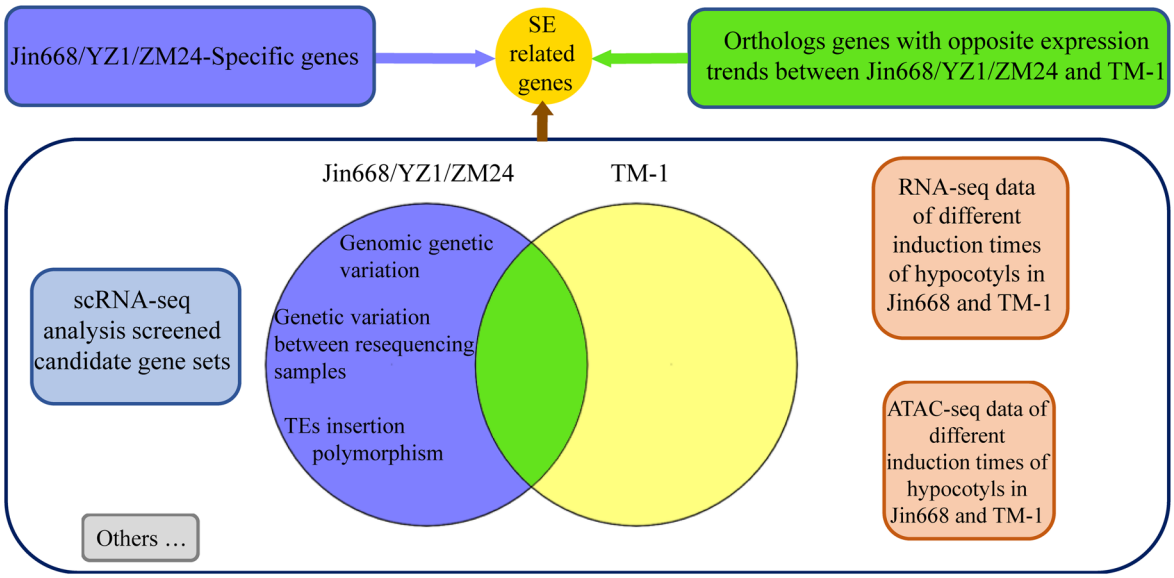
Extended Data Fig. 5 | A comparative genome analysis of the Jin668 versus TM-1, YZ1, and ZM24. a, Genome-wide syntenic relationships among At and Dt subgenomes in four Upland cotton accessions relative to the A-genome-like Ga (A_2 genome) and D-genome-like Gr (D_5 genome). **b**, The distribution of variation density between Jin668 versus TM-1, YZ1, and ZM24 genomes. **c**, The distribution of SE-related genes and gene density, as well as genomic variations from Fig.1, on the chromosomes of the Jin668 genome. The circos plot from outside to inside shows: chromosome, SE-related genes, gene density in Jin668, SNPs, InDels, PAVs, inversions and translocation density between Jin668 with TM-1. **d**, The correlation between SNPs, InDels, PAVs, inversions and translocations with SE-related genes. The significance was tested using the overlapPermTest function in regioneR (https://www.bioconductor.org/packages/devel/bioc/html/regioneR.html), which calls permTest with the appropriate parameters to perform the permutation test. SE, somatic embryogenesis.

https://doi.org/10.1038/s41588-025-02258-3



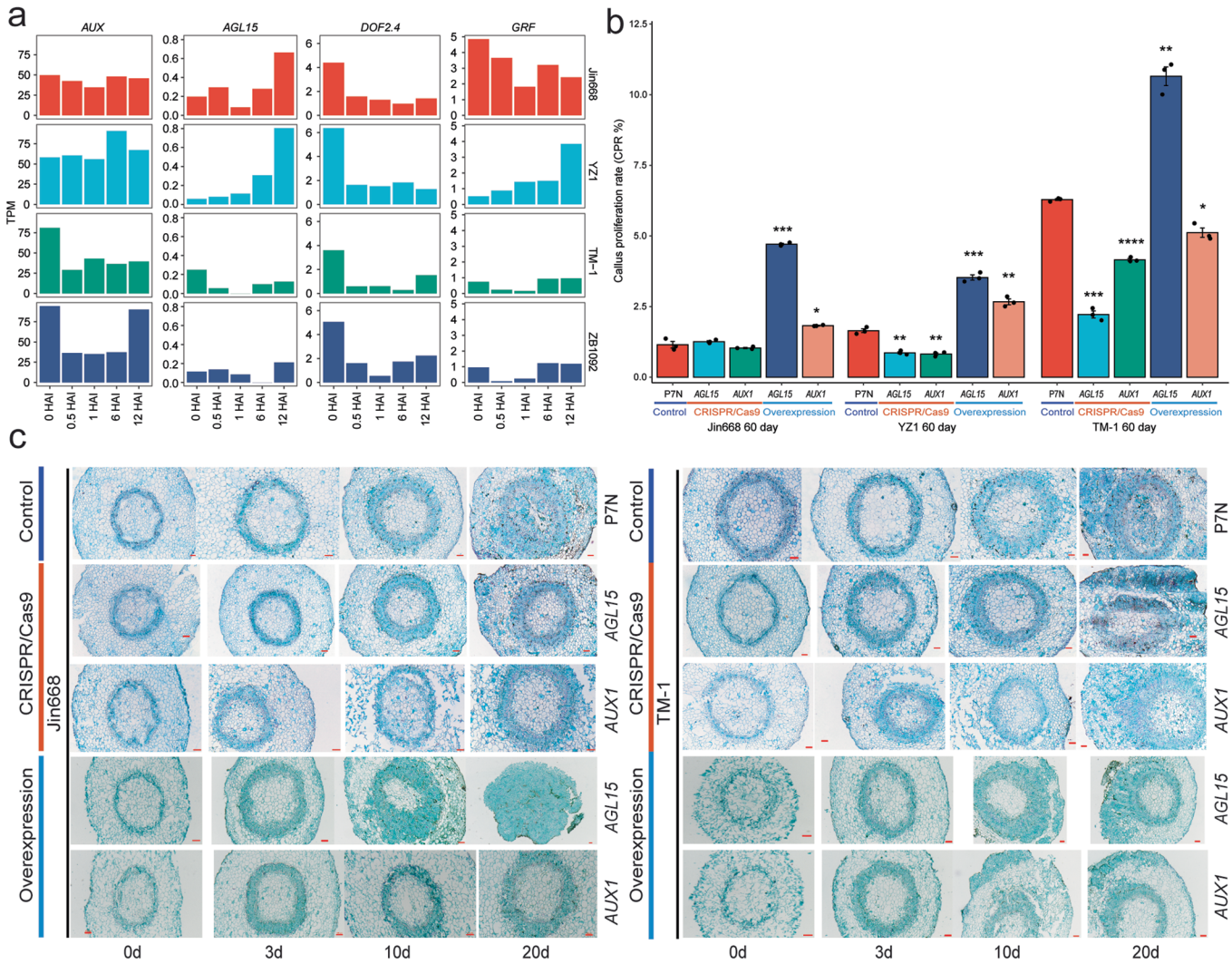
Extended Data Fig. 6 | Sample preparation, sequencing, and quality evaluation of ATAC library. a, Schematic outline of genome-wide ATAC−seq and RNA-seq assays and time points of sample collection with two independent biological replicates, as well as the roadmap of an integration analysis. HAI, hours after inoculation. **b**, The principal component plots of ATAC−seq data sequenced from Jin668, YZ1, TM-1 and ZB1092, respectively. Color code is shown. Each dot represents one sample. **c**, The genome-wide distribution of ATAC−seq peaks for Jin668, YZ1, TM-1 and ZB1092. Window size: TSS ± 3.0 Kb. **d**, The distribution of SPOT values for all samples from Jin668, YZ1, TM-1 and ZB1092, respectively. **e**, The distribution of FRiP values for all samples from Jin668, YZ1, TM-1 and ZB1092, respectively. **f**, ATAC−seq profiling in Jin668, YZ1, TM-1 and ZB1092. Left, the number statistics of chromatin accessibility peaks on each chromosome. Middle, the bar chart shows the total number of peaks detected at the corresponding induction time point. Right, the pie chart illustrates the percentage distribution of peaks across various genomic regions.

https://doi.org/10.1038/s41588-025-02258-3



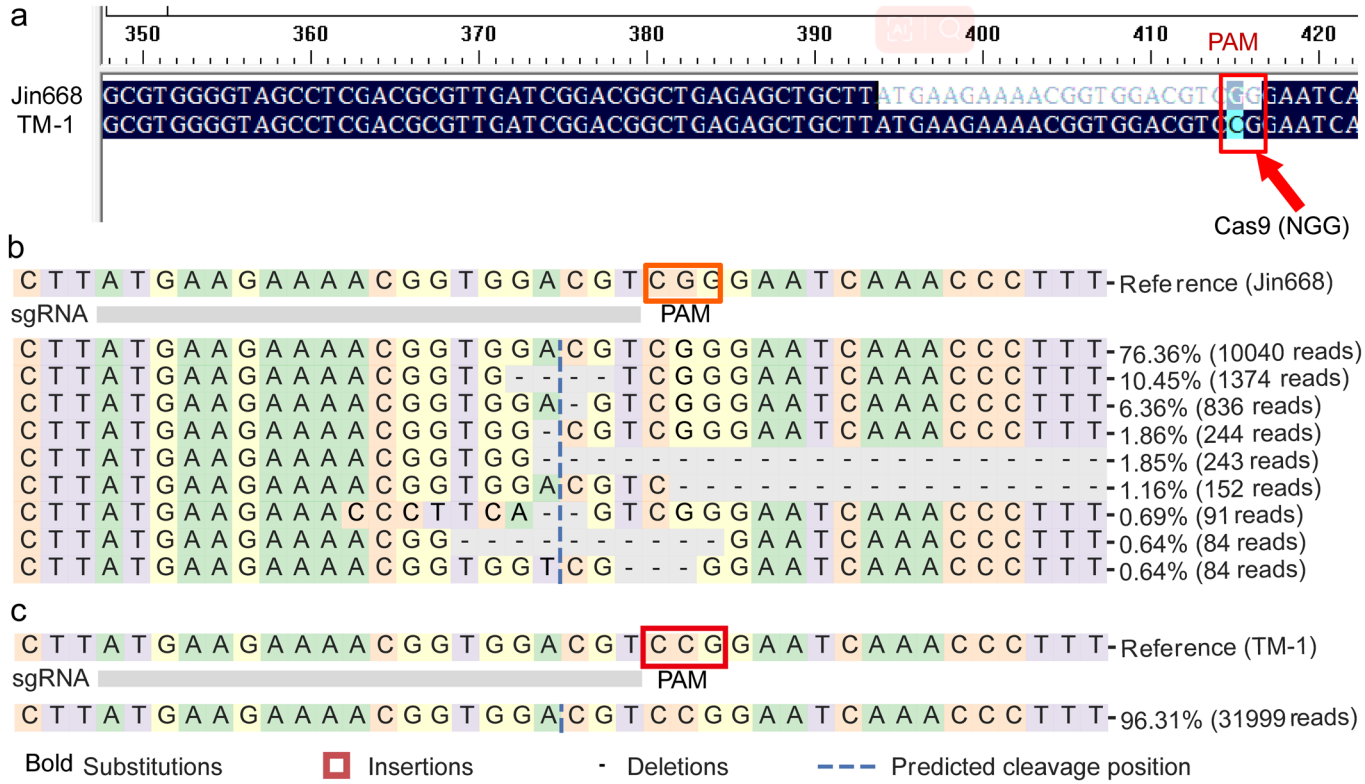
Extended Data Fig. 7 | Identification of SE-related genes. Identification of SE-related genes based on genome information, chromatin accessibility and gene expression change trend at different time points of SE stage, and other external data (published literature and data generated in our laboratory earlier, etc.).

https://doi.org/10.1038/s41588-025-02258-3



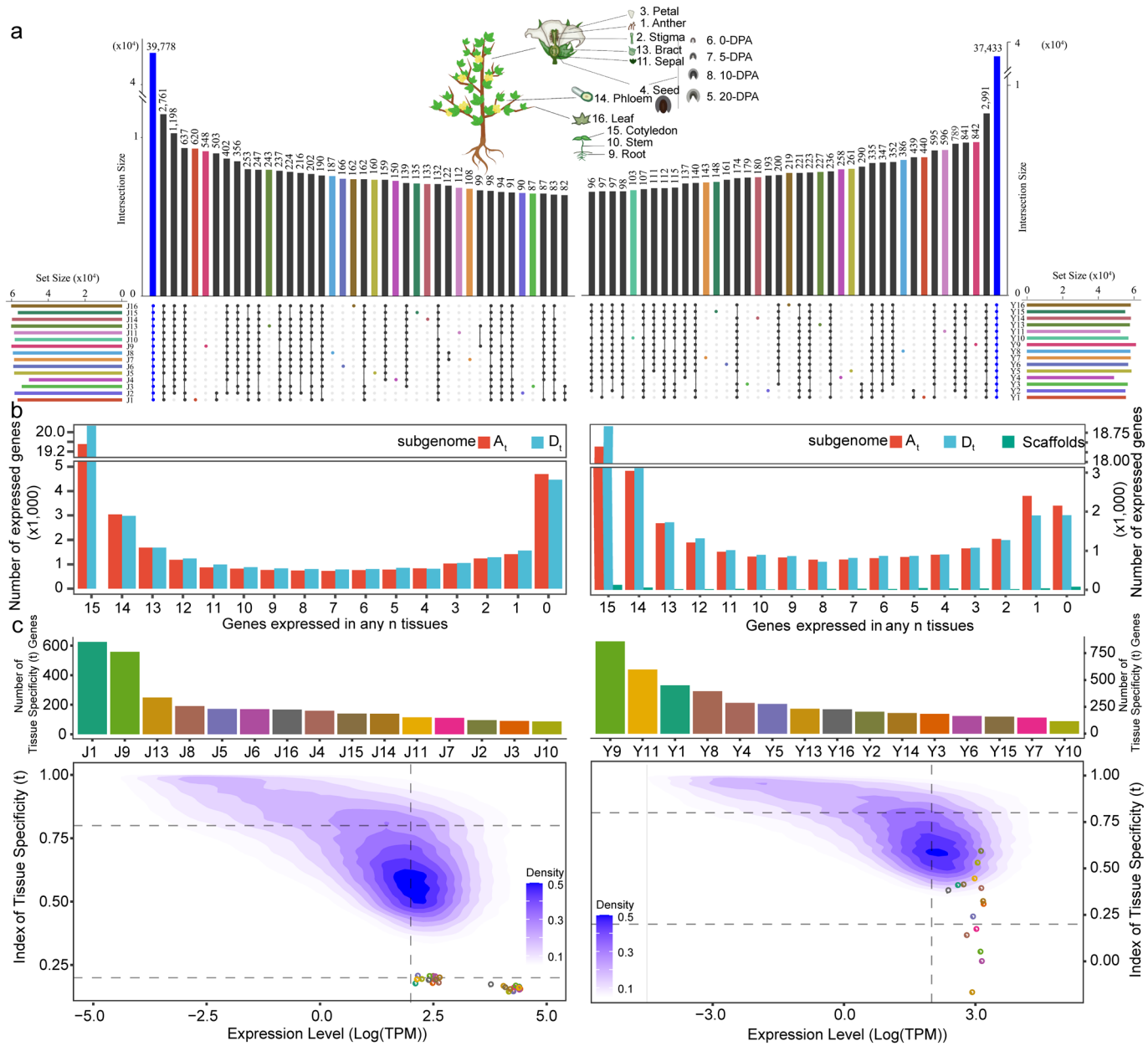
Extended Data Fig. 8 | *AGL* promotes cotton regeneration. a, Transcription and epigenetic tracks for *AGL*. Transcription data shown as mean ± s.d. of 3 biological replicates. **b**, The CPR of CRISPR edited, overexpression and control lines at 60 days post-induction. Data are represented as mean ± s.d. Differences between groups were evaluated by two-sided Student's *t*-test, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. n = 3 independent biological replicates. The P7N was used as control. **c**, The paraffin sections of CRISPR edited, overexpression and control lines for Jin668 and TM-1 at different days post-induction. Scale bars represent 100 μm.

https://doi.org/10.1038/s41588-025-02258-3



Extended Data Fig. 9 | Genetic transformation experiments demonstrate that the Jin668 genome enhances the accuracy of sgRNA design for CRISPR gene editing. a, The sequence homology of a putative sgRNA (present in Jin668 but absent in TM-1) between Jin668 and TM-1. **b**,**c**, The mutational landscape of the genomic regions corresponding to targets sgRNA within the callus of Jin668 (**b**) and TM-1 (**c**) after edited using CRISPR/Cas9. The percentage signifies the ratio of reads corresponding to this mutation type to the total read count.

https://doi.org/10.1038/s41588-025-02258-3



Extended Data Fig. 10 | Gene expression dynamics across 15 tissues in Jin668 and YZ1. a, The profile of genes expressed in all 15 tissues of Jin668 (left) and YZ1 (right). **b**, Statistics on the number of genes co-expressed in multiple tissues of Jin668 (left) and YZ1 (right). **c**, Statistics on the number of genes specifically expressed in each tissue of Jin668 (left) and YZ1 (right). Each tissue is represented by the number corresponding to that shown in the graphic illustration in **a**.

nature portfolio

Corresponding author(s): Shuangxia Jin

Last updated by author(s): Mar 26, 2025

Reporting Summary

Nature Portfolio 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 Portfolio 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	No software was used during data collection.
Data analysis	All open source softwares or piplines (versions) used in this study were cited in the Method. These include jellyfish (v2.3.0), GenomeScope (v2.0), Guppy (v6.3.4), ccs (v6.0), NextDenovo (v2.5.0), NextPolish (v1.4.1), Hifiasm (v0.18.1), Polish (v.1.23), Bionano Solve Pipeline (v3.5.1_01142020; https://bionanogenomics.com/support/software-downloads/), 3D-DNA (v180114), Juicebox Assembly Tools (JBAT v1.8.8), Filtlong (v0.2.1), BWA-MEM (v0.7.17), DeepVariant (v1.4.0), Sniffles (v2.0.2), BCFtools (v1.8), Merfin (v1.1), Winnowmap2 (v2.03), Medaka (v1.6.0), Racon (v1.6.0), Teloclip (https://github.com/Adamtaranto/teloclip), pyTanFinder (https://github.com/Kirovez/pyTanFinder), BLAST (v2.7.1+), Trim Glore (v0.6.7), BUSCO (v5.2), Epic2 (https://github.com/biocore-ntnu/epic2), RepeatModeler (v2.0.4), RepeatMasker (v4.0.7), Repbase (Repbase21.08), EDTA (v2.0.0), Augustus (v3.3.1), Genscan (v3.1), GeneID (v1.4), GlimmerHMM (v1.2), GeneMarkS-T (v4), SNAP (v2006-07-28), GeMoMa (v1.9), GeneWise (v2.4.1), HISAT2 (v2.2.1), StringTie (v2.1.4), EVidenceModeler (EVM; v1.1.1), InterProScan (v5.60-92.0), tRNAscan-SE (v1.23), Infernal (v1.1.2), RNAmmer (v1.2), HiC-Pro pipeline (v2.11.14), HiCExplorer (v3.7.2), Trimmomatic (v0.39), minimap2 (v2.16-r922), OrthoMCL (v2.0.9), mafft (v7.505), IQ-TREE (v1.6.12), jcvi (https://github.com/tanghaibao/jcvi), circos (v0.69-4), SyRI (v1.6), SnpEff (v4.3), Genome Analysis Toolkit (GATKv4.1), pbmm2 (v1.16.0), Samblaster (v0.1.26), MACS2 (v2.2.7.1), DiffBind (v2.14.0), CRISPOR (v3.1), R (v4.0.0), Django (v4.2.5), picard (v2.23.9), deepTools (v3.5.0), bedtools (v2.27), NucFreq (https://github.com/mrvollger/NucFreq), TandemTools (https://github.com/ablab/TandemTools), DANTE-Protein Domain Finder (v0.2.5), PATHd8 (v1.0), GenMap (v1.3.0), karyoploteR (v1.8.4), scanPAV (https://github.com/wtsi-hpag/scanPAV), Samtools (v1.6), Cytoscape (v3.10.3), KOBAS (v3.0), fastp (v0.23.1), HOMER (v4.11.1), Lima (v2.9.0), isoseq3 (v4.0.0), Pilon (v1.24), TRF (v4.0.9), ribotin (https://github.com/maickrau/ribotin), BBMap (v35.85), Merqury (v1.3), Meryl (v1.3), Trinity (v2.11.0), PASA (v2.4.1), TransDecoder (v5.7.1), GMAP (v2020-03-12), KEGG Automatic Annotation Server (v2.1), seqtk (https://github.com/lh3/seqtk), BISER (https://github.com/0xTCG/biser), Primrose (https://github.com/PacificBiosciences/primrose). Some customized Python (v2.7.15 & v3.11.4) scripts were used to process the data generated by each software, which parameters

were described in Methods section. All original codes have been deposited at Github (https://github.com/tiramisutes/T2T-Cotton-Genomes) and is publicly available as of the date of publication.

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 Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The Jin668 and YZ1 assemblies and annotation data are available at T2TCotton-Hub (http://jinlab.hzau.edu.cn/T2TCottonHub/). The raw sequencing data used for de novo whole-genome assembly of Jin668 and YZ1 are available in NCBI under the BioProject: PRJNA874817 and PRJNA960814, respectively. The RNA-seq of Jin668 and YZ1 are available in NCBI under the BioProject: PRJNA874819 and PRJNA960820, respectively. The ATAC-seq data of Jin668 is available in NCBI under the BioProject PRJNA960832. The ATAC-seq and RNA-seq data of TM-1 during somatic embryogenesis are available in NCBI under the BioProject: PRJNA960828 and PRJNA960825, respectively. The ATAC-seq and RNA-seq data of YZ1 during somatic embryogenesis are available in NCBI under the BioProject: PRJNA1059614 and PRJNA1059613, respectively. The ATAC-seq and RNA-seq data of ZB1092 during somatic embryogenesis are available in NCBI under the BioProject: PRJNA1059611 and PRJNA1059609, respectively. The ChIP-seq data for Jin668 and YZ1 are uploaded to BioProject PRJNA1079680 and PRJNA1079682, respectively. The seeds of Jin668 and YZ1 used in this study are available from the corresponding author upon request. The reference genome assembly and annotation files of TM-1 (v3.1) used in this study were downloaded from https://phytozome-next.jgi.doe.gov/ and can also be downloaded from http://jinlab.hzau.edu.cn/T2TCottonHub/. In addition, we collected all available SE-related motifs from Plant Transcription Factor Database (PlantTFDB; v5.0; https://planttfdb.gao-lab.org/). Further details on data accessibility are outlined in the Supplementary Materials and Methods.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Not suitable for our study.
Reporting on race, ethnicity, or other socially relevant groupings	Not suitable for our study.
Population characteristics	Not suitable for our study.
Recruitment	Not suitable for our study.
Ethics oversight	Not suitable for our study.

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

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	For RNA-Seq and ATAC-seq analysis experiments of different tissues of Jin668, YZ1 and TM-1 and used two biological replicates. No statistical methods were used to predetermine sample sizes.
Data exclusions	All sequences data in this study, we only excluded sequences that were of low quality. For genome assembly, we only excluded the redundant sequences that were caused because of heterozygosity.
Replication	There is various approaches were used to evaluate the completeness of assemblies genome. The quality of the data was high and all replicates reported in the manuscript are biological replicates. For all biological experiments (ATAC-seq, RNA-seq and FISH), at least two independent replications were performed. All attempts at replication were successful.
Randomization	The plants used for sequence were all grown in the same conditions and were placed randomly in the growth room.

Blinding	Data collection and analysis were performed in a blinded manner.

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
☒	☐ Clinical data
☒	☐ Dual use research of concern
☐	☒ Plants

Methods

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

Antibodies

Antibodies used	CENH3 antibodies. The CENH3 antibody is a rabbit polyclonal against the peptide MSRTKHTAAKKP. This antibody is not a fully developed commercial product; it was generated by immunizing rabbits and subsequently purified.
Validation	The rabbit polyclonal antibody against CENH3 was validated to be available according to a previous report (Han, J. et al. Rapid proliferation and nucleolar organizer targeting centromeric retrotransposons in cotton. Plant J. 88, 992–1005 (2016).).

Plants

Seed stocks	The seeds of Jin668 and YZ1 used in this study are available from the corresponding author upon request.
Novel plant genotypes	Jin668 was artificially domesticated through a Successive Regeneration Acclimation (SRA) strategy (Li, J. et al. Multi-omics analyses reveal epigenomics basis for cotton somatic embryogenesis through successive regeneration acclimation process. Plant Biotechnology Journal 17, 435-450 (2019).).
Authentication	Many studies have reported the use of Jin668 and YZ1 for functional genome research in cotton. Such as: Wang, G., Wang, F., Xu, Z. et al. Precise fine-turning of GhTFL1 by base editing tools defines ideal cotton plant architecture. Genome Biol 25, 59 (2024). https://doi.org/10.1186/s13059-024-03189-8. He, X., Wang, T., Xu, Z. et al. The cotton HD-Zip transcription factor GhHB12 regulates flowering time and plant architecture via the GhmiR157-GhSPL pathway. Commun Biol 1, 229 (2018). https://doi.org/10.1038/s42003-018-0234-0.

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.*	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1079680/ https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1079682/
Files in database submission	PRJNA1079680, PRJNA1079682
Genome browser session (e.g. UCSC)	No longer applicable.

Methodology

Replicates	One biological replicate was set.
Sequencing depth	Jin668 (Depth 10×; Total reads: 37,208,070; mapped reads: 35,980,822; 150 bp pared -end reads.) YZ1 (Depth 10×; Total reads: 32,303,777; mapped reads: 19,247,252; 150 bp pared -end reads.)
Antibodies	anti-CENH3

Peak calling parameters	Enrichment leve of CENH3 for each base was obtained using bamCompare in the Deeptools package (v3.5.0) with the parameters of '-binSize 1--numberOfProcessors 40 --operation ratio --outFileFormat bedgraph'. Average enrichment of each 1 kb-bin of the genome was then calculated. The bins that enrichment levels greater than 5 were retained, which with a distance interval less than 1Mb were merged. The final centromeric regions were determined by visual inspection of the distribution of CENH3 ChIP-seq peaks.
Data quality	All 26 centromeres in Jin668 and YZ1 were successfully identified.
Software	Enrichment level of CENH3 for each base was obtained using bamCompare in the Deeptools package (v3.5.0). In addition, Epic2 (https://github.com/biocore-ntnu/epic2) with the CENH3 ChIP-seq alignment as treatment, input samples read alignment as control, MAPQ (mapping quality) as 20, effective genome fraction, bin size as 5000, and gap size as 0 was employed to call peaks.