

Genetic architecture of sugarcane traits in a polyploid genomics framework

<https://doi.org/10.1038/s41586-026-10576-7>

Received: 9 April 2025

Accepted: 21 April 2026

Published online: 27 May 2026

Open access

 Check for updates

Jungang Wang (王健刚)^{1,11}, Xiaofeng Li (李晓峰)^{2,11}, Yibin Wang (王毅斌)^{2,11}, Jishan Lin (林继山)^{1,11}, Shuai Chen (陈帅)^{2,11}, Hongbo Liu (刘洪博)^{3,11}, Xiao Chen (陈晓)^{2,11}, Kun Chai (柴琨)^{2,4,5}, Aoqian Dong (董澳千)⁴, Tingting Zhao (赵婷婷)¹, Cuilian Feng (冯翠莲)¹, Ruijie Wu (吴蕊洁)¹, Ping Zhao (赵萍)², Yaodong Zheng (郑曜东)⁴, Zhongqiang Xia (夏中强)², Shengcheng Zhang (张晟诚)², Yi Liu (刘怡)², Shenyang Qu (屈伸洋)², Ziqi Ye (叶子奇)², Yuhao Song (宋雨函)², Qingyuan Deng (邓清源)², Xiaofei Zeng (曾筱菲)², Guangrun Yu (余光润)², Ran Kong (孔冉)⁶, Baoqing Zhang (张保青)⁷, Wei Zhang (张伟)⁸, Peifang Zhao (赵培方)³, Jun Mao (毛钧)³, Xin Lu (陆鑫)³, Haifeng Jia (贾海凤)², Xueting Zhao (赵雪婷)^{1,9}, Qianqian Zhang (张倩倩)^{1,9}, Shuzhen Zhang (张树珍)^{1,9}, Wenwei Cai (蔡文伟)^{1,9}, Dongao Huo (霍冬敖)^{1,9}, Ling Li (李玲)^{1,9}, Yuqing Gong (宫雨晴)⁴, Shiqiang Huang (黄世强)⁴, Yongji Huang (黄永吉)¹⁰, Zehuai Yu (余则怀)⁵, Zuhu Deng (邓祖湖)⁴, Baoshan Chen (陈保善)⁵, Yuebin Zhang (张跃彬)³, Muqing Zhang (张木清)^{5,10}, Ray Ming (明瑞光)^{4,10} & Xingtian Zhang (张兴坦)^{2,10}

Sugarcane (*Saccharum* spp.) is a vital sugar and bioenergy crop with an exceptionally complex polyploid genome (10–12 sets of chromosomes). This complexity resulted from nobilization—a historical breeding process involving interspecific hybridization and repeated backcrossing¹. However, the extreme ploidy has long impeded efforts to elucidate the genetic basis of its considerable sucrose-storing capacity. Here we present a fully phased genome assembly of the foundational cultivar POJ2878, achieved using a Pore-C-based assembly algorithm. This assembly resolved 118 chromosomes, revealing extensive subgenome recombination and non-homologous chromosomal rearrangements. Using identity-by-descent and allele-specific expression profiling, we identified breeder-favoured haplotypes, including a *SUS2* haplotype with enhanced sucrose content. Resequencing of 981 *Saccharum* accessions traced POJ2878's pervasive contribution to modern cultivars and identified key domestication and improvement sweeps. Genes under selection include *CBL1* for cold tolerance, *TIP1* for cell size regulation and *TBI* for tillering control. A genome-wide association study tailored for polyploid genomes resolved loci associated with parenchyma cell size and sucrose storage capacity, including the functionally validated sucrose transporter *Saccharum* hybrid *SUT2*. These findings clarify the genetic architecture underlying sugarcane's biomass productivity and sugar yield, offering a genomic foundation for accelerating improvement in sugarcane and other polyploid crops critical for global food and bioenergy security.

It is hard to imagine a world without readily available sweet treats, but that was the reality before sugarcane was domesticated. Early humans depended on scarce and seasonal sources of sweetness, such as honey, which were considered luxuries. The domestication of sugarcane revolutionized sugar production, transforming it into a global dietary staple. However, this transformation came with profound costs: widespread availability and overconsumption of sugar have contributed to a global health crisis, and the expansion of sugar production was deeply intertwined with the transatlantic slave trade, which reshaped human migration, cultures and economies at the global scale.

Sugarcane is the world's leading sugar crop, renowned for its exceptional ability to accumulate sucrose in stalk parenchyma cells.

However, the genetic mechanisms underlying this trait remain largely unexplored. Evidence in other crops suggests a strong association between parenchyma cell size and sugar accumulation. For example, in sugar beet (*Beta vulgaris*), *BZRI* promotes parenchyma cell expansion, enhancing sugar loading efficiency², and, in strawberry (*Fragaria vesca*), the tonoplast transporter *MATE51* coordinates sugar transport with cell size³. In sugarcane, modifications in parenchyma cell wall elasticity allow cells to withstand high turgor pressure from sucrose accumulation⁴. Nevertheless, the specific genes and regulatory networks governing these processes in sugarcane remain unclear, limiting our understanding of its exceptional productivity as a sugar crop.

A list of affiliations appears at the end of the paper.

Modern sugarcane improvement has been shaped by hybridization between *Saccharum officinarum* and *Saccharum spontaneum*, followed by repeated backcrossing with *S. officinarum*—a process known as nobilization¹. Together with evidence for an unknown *Saccharum* contributor in early hybridization⁵, these processes have merged divergent *Saccharum* lineages, resulting in extreme polyploidy and genomic complexity. The evolutionary history has left genome-wide signatures of introgression, identity-by-descent (IBD) from founder cultivars, and selective sweeps associated with domestication and modern breeding. Although challenging, dissecting these signatures using haplotype-resolved references and allele-aware analyses offers a promising path to unravelling the considerable adaptability and productivity of sugarcane.

Advances in long-read sequencing technologies, such as PacBio HiFi and BioNano, have proven to be effective in resolving polyploid genomes, as demonstrated by recent assemblies for hybrid cultivars^{6–9}. However, gaps remain in achieving full haplotype-level resolution, understanding allele-specific expression and deciphering the molecular mechanisms that drive high sugar capacity. Short-read sequencing, despite aiding high-throughput chromosome conformation capture (Hi-C) and RNA-sequencing (RNA-seq) applications, struggles to resolve homologous or homoeologous sequences, leading to errors in haplotype phasing and allelic expression analyses^{10,11}.

Emerging long-read technologies—Oxford Nanopore Technologies Ultra-Long (ONT-UL) reads, Pore-C sequencing¹² and multiplexed arrays isoform sequencing (MAS-ISO-seq)¹³—offer possibilities for accurate genome assembly and allelic resolution. Yet, their application to sugarcane's ultra-complex genome has been limited by the absence of tailored bioinformatics tools. Here we integrate C-Phasing (a haplotype-resolved assembly algorithm¹⁴), KMERIA (a polyploid genome-wide association study (GWAS) tool¹⁵) and Allele-Express (an allele-specific expression tool) to generate a fully phased genome for the hybrid cultivar POJ2878. By resequencing 981 *Saccharum* accessions, we traced sugarcane's domestication and improvement, identifying key genetic shifts and loci associated with sugar storage. This work illuminates sugarcane's molecular evolution and establishes a robust framework for polyploid genomics, positioned to accelerate breeding in sugarcane and other crops with complex genomes.

A fully phased genome assembly

The POJ2878 cultivar, known as the 'King of sugarcane', was selected for de novo genome sequencing and assembly due to its role as a key genetic donor for modern sugarcane cultivars. A high-quality genome assembly was generated using advanced sequencing technologies, including 678.52 Gb of MGI-seq short reads, 405.42 Gb of PacBio HiFi reads and 148.64 Gb of ONT-UL reads with N50 lengths over 100 kb (Supplementary Table 1). The average read lengths are 15.84 kb (PacBio) and 161.94 kb (ONT-UL) (Supplementary Fig. 2). Pore-C technology, an ONT-based method capturing high-order chromatin 3D structure, produced 517.13 Gb of reads with 2.60 kb average length and 4.15 kb N50. Notably, 15.15% of Pore-C reads (22.31 million out of 147.25 million) contained at least three chromatin concatemers (Supplementary Table 2 and Supplementary Fig. 3). Together, these datasets provided 376× coverage of the estimated 10-Gb genome. Hifiasm was used with the HiFi and ONT-UL reads to yield a 10.37-Gb contig-level assembly with a 15.60-Mb N50 (Extended Data Table 1 and Supplementary Table 3).

The C-Phasing algorithm¹⁴, integrating Pore-C long-range contact information, was used to achieve a fully haplotype-resolved, chromosomal-scale genome assembly (Fig. 1a). Pore-C technology substantially improved the proportion of uniquely aligned reads (75.30% versus 39.81%), and the rate of valid interaction pairs (36.44% versus 29.69%; Supplementary Table 2), producing a clearer chromatin interaction heat map than Hi-C technology (Supplementary Fig. 4).

Using a hyperpartition algorithm, C-Phasing resolved 118 groups, each representing one chromosome (Extended Data Table 1), matching the karyotype determined using fluorescence in situ hybridization (FISH) (Fig. 1b). After manual correction, the final assembly comprises 10 homologous/homoeologous groups, each containing an average of 12 chromosomes (Fig. 1c).

The genome assembly quality was assessed across three dimensions: completeness, contiguity and accuracy (Supplementary Note 1 and Supplementary Table 4). Notably, the assembly demonstrated a high level of contiguity, achieving a contig N50 of 15.60 Mb, along with extremely low error rates, including 2.02% switched and 0.82% collapsed genomic regions (Supplementary Fig. 5). Chromatin contacts showed stronger intrasubgenome interactions than intersubgenome interactions (Fig. 1d). These results highlight the high quality of the POJ2878 genome assembly, with most evaluation metrics meeting or exceeding those of previously published hybrid sugarcane genomes.

Recombination and rearrangement

The genomic components of POJ2878 was analysed using resequencing data from 10 *S. officinarum* and 10 *S. spontaneum* individuals (Methods). This analysis revealed that 80.5% (8.35 Gb) of the genome originated from *S. officinarum* and 19.5% (2.02 Gb) from *S. spontaneum* (Supplementary Fig. 6). We further identified marked recombination events between the *S. officinarum* and *S. spontaneum* subgenomes, with recombined regions constituting at least 10% of their length. This classified the genome into 92 *S. officinarum*-derived, 16 *S. spontaneum*-derived and 10 recombinant chromosomes (Extended Data Fig. 1a and Supplementary Table 5).

We identified 13.23 Mb of small-scale homoeologous exchanges distributed across 49 *S. officinarum*-derived and 2 *S. spontaneum*-derived chromosomes, representing 0.13% of the genome (Supplementary Table 5) and encompassing 320 protein-coding genes (11 *S. officinarum* and 309 *S. spontaneum* derived). Functional analysis showed that *S. officinarum*-derived genes were enriched in processes such as stimulus response and intracellular signal transduction, whereas *S. spontaneum*-derived genes were enriched in post-embryonic development and system development (Fisher's exact test, $P < 0.05$; Supplementary Fig. 7).

To investigate chromosomal rearrangements, we generated telomere-to-telomere (T2T) monoploid genome assemblies for the progenitor species *S. officinarum* XZ and *S. spontaneum* 82-114 (Supplementary Note 2). Synteny analysis revealed high overall collinearity and frequent non-homologous chromosomal rearrangements in POJ2878 (Extended Data Fig. 1), including confirmed chromosome fission and fusion events in *S. spontaneum*-derived subgenome¹⁶ and four large-scale non-homologous rearrangements in the *S. officinarum*-derived subgenome. Notably, chromosome 5g1 resulted from a crossover, with its first 42 Mb derived from the *S. spontaneum* subgenome and the rest from the *S. officinarum*-derived subgenome. This structure was independently validated using haplotype-specific probes in oligo-FISH experiments (Fig. 1b), confirming the recombination event and assembly accuracy.

Telomeric analysis revealed that all 118 chromosomes had at least one telomere annotated, with 73 (61.86%) chromosomes containing telomeres at both ends (Extended Data Fig. 2 and Supplementary Table 6). Centromeric analysis identified 14 telocentric, 6 acrocentric, 33 submetacentric and 65 metacentric regions with averaging 2.22 Mb. Notably, 40% of the 10 *S. spontaneum*–*S. officinarum* recombined chromosomes were telocentric and acrocentric.

Haplotype-specific gene expression

The POJ2878 genome is composed of 69.24% (7.18/10.37) transposable elements with long terminal repeat (LTR) retrotransposons being

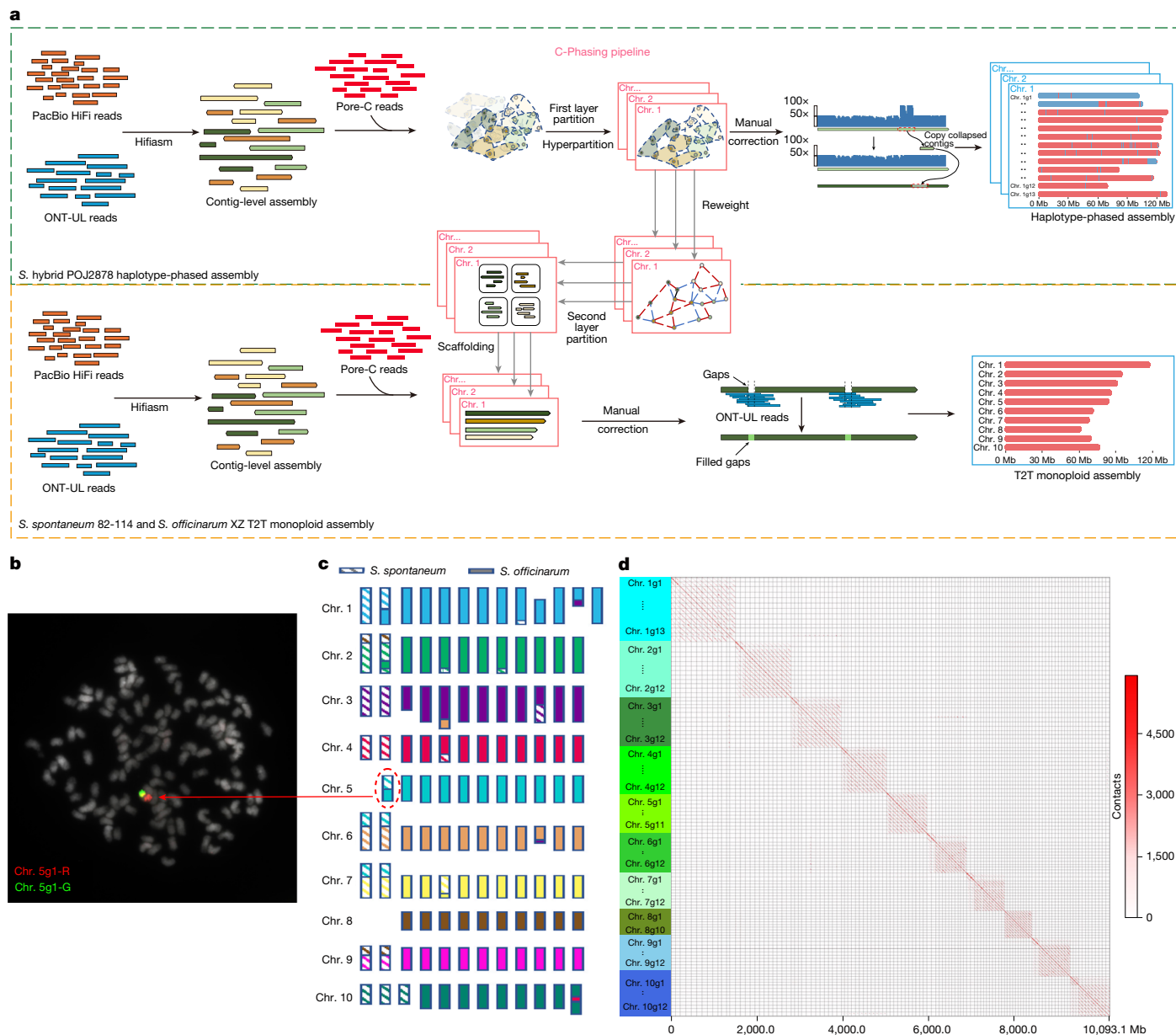


Fig. 1 | Genome assembly, subgenome identification and chromosomal rearrangement. **a**, Schematic of the pipeline used to generate a haplotype-resolved genome assembly for the *Saccharum* hybrid cultivar POJ2878 and T2T monoploid genome assemblies for the progenitor species, *S. spontaneum* 82-114 and *S. officinarum* XZ. The genomes were sequenced using the PacBio HiFi, ONT-UL and Pore-C sequencing technologies. For POJ2878, phased unitigs (p_utg) assembled by hifiasm were partitioned into homologous or homoeologous groups and haplotypes using C-Phasing with Pore-C contacts, and collapsed regions were recovered based on sequencing depth to produce a fully haplotype-resolved assembly. For the progenitor T2T monoploid assemblies, hifiasm contigs (p_ctg) were deduplicated, scaffolded into monoploid chromosomes using C-Phasing and gap-filled with ONT-UL reads. **b**, Oligonucleotide painting

(oligo-FISH) of chromosome (chr.) 5g1 using two probe sets (chromosome 5g1-R and chromosome 5g1-G) labelling *S. spontaneum*-derived (red; 1.56–42 Mb) and *S. officinarum*-derived (green; 42–81.72 Mb) segments, respectively. The image is representative of one metaphase cell from one POJ2878 plant. **c**, A proposed model of chromosomal rearrangement based on subgenome assignments and synteny analysis. Different colours represent distinct homologous or homoeologous groups; the shaded rectangles indicate *S. spontaneum*-derived sequences and the filled rectangles represent *S. officinarum*-derived sequences. **d**, Raw chromatin interactions among the 118 chromosomes, highlighting intrachromosomal and interchromosomal interactions.

the predominant class in both subgenomes (52.04% in the *S. spontaneum* subgenome and 59.05% in the *S. officinarum* subgenome; Supplementary Table 7 and Supplementary Fig. 8a). Both values exceed the proportions observed in the progenitor genomes, with 41.36% in *S. spontaneum* 82-441 and 49.48% in *S. officinarum* XZ. This elevated LTR content reflects lineage-specific LTR bursts, occurring around 0.16 million years ago in the *S. officinarum*-derived subgenome and around 0.2 million years ago in the *S. spontaneum* subgenome

(Supplementary Fig. 8b), predating POJ2878's hybridization events less than 120 years ago. These results suggest that the actual progenitors used in producing POJ2878 differ from those analysed here.

A comprehensive annotation pipeline initially identified 398,858 protein-coding genes (including allelic genes), followed by classification of genes with defined alleles, paralogous genes and haplotype-specific genes (Methods and Extended Data Fig. 3). This produced a

high-quality allele-defined annotation with 98.6% BUSCO completeness (Supplementary Table 8), encompassing 35,082 and 28,851 genes in the *S. officinarum*- and *S. spontaneum*-derived subgenomes, respectively (Extended Data Table 1 and Supplementary Table 9). These allele-defined genes represent instances in which multiple allelic variants are present within the genome, with each allele differing by at least one nucleotide variant in coding sequences. A total of 28,523 homoeologous gene pairs were identified between the two subgenomes.

Quantifying haplotype-specific gene expression in polyploid genomes is challenging due to the inability of short RNA-seq reads to accurately distinguish highly similar haplotypes. To address this, we used MAS-ISO-seq¹³, which captures full-length transcripts by ligating multiple cDNA fragments into single molecules, and developed Allele-Express, specifically designed to improve allele-specific expression analysis in polyploids (Extended Data Fig. 4). Benchmarking showed that Allele-Express achieved higher accuracy compared with RNA-seq methods ($R^2 = 0.95$ versus $R^2 = 0.70$; Supplementary Note 3).

Using MAS-ISO-seq and Allele-Express, we sequenced transcripts from mature leaf and stem tissues, generating 63.88 million and 56.23 million full-length transcripts, respectively (Supplementary Table 10). Each replicate yielded an average of 30.75 Gb (leaf, 1,444 bp mean length) and 27.83 Gb (stem, 1,476 bp mean length), with replicates showing high consistency (Pearson correlation > 0.98; Supplementary Fig. 9). A saturation curve showed that transcript discovery plateaued beyond 18 Gb (30× transcriptome coverage; estimated transcriptome size: 616.78 Mb; Supplementary Fig. 10). We identified 6,994 allele-specific highly expressed genes (ASHEGs) in mature stem and 7,325 in mature leaf samples, where one allele was significantly more highly expressed than its counterparts (G -test, $P < 0.05$). ASHEGs were unevenly distributed across chromosomes, with significant enrichment observed on five chromosomes in stem tissue and nine in leaf tissue (binomial test with Bonferroni correction, adjusted $P < 0.01$; Supplementary Fig. 11a,b). This enrichment followed tissue-specific chromosomal activation patterns, with seven out of ten enriched chromosomes originating from the *S. spontaneum* subgenome and only three from the *S. officinarum*-derived subgenome. The *S. spontaneum* subgenome displayed significantly higher ASHEG density than the *S. officinarum*-derived subgenome in both stem (0.024 versus 0.016, Mann–Whitney U -test, $P < 0.001$) and leaf (0.026 versus 0.017, $P < 0.001$) tissues (Supplementary Fig. 11c,d), despite comprising only approximately 20% of the POJ2878 genome. This indicates a disproportionate contribution of *S. spontaneum*-derived alleles to allele-specific high expression.

Functional enrichment analysis revealed that these ASHEGs were associated with critical biological processes and cellular components (Supplementary Figs. 12 and 13). In leaf, ten ASHEGs were enriched in cotyledon morphogenesis, and seven in the positive regulation of auxin-mediated signalling pathways, including genes such as *PINI* (auxin transport¹⁷) and CSA-like transcription factors (regulating sugar content¹⁸; Supplementary Fig. 14). In stem tissue, ASHEGs were enriched in storage vacuoles (18 genes) and COPI-coated vesicles (9 genes; Supplementary Fig. 15), with examples including the probable aquaporins *TIP2* and *TIP3* (cell turgor pressure and cell expansion¹⁹) and *beta-COP1* (light signalling regulation²⁰).

Resequencing and population structure

We resequenced 981 *Saccharum* samples, including 78 *S. officinarum*, 290 *S. spontaneum* and 613 hybrid sugarcane cultivars, representing 19 major sugarcane-producing regions worldwide (Fig. 2a–c). This generated 125.58 Tb of short-read data with an average depth of 16.57× per sample (Supplementary Table 11). Analysis of uniquely aligned reads to the POJ2878 reference genome yielded 11.43 million single-nucleotide polymorphisms (SNPs) and 7.36 million insertion–deletions (indels).

Phylogenetic analysis using high-quality SNPs revealed three distinct clades (*S. spontaneum*, *S. officinarum* and hybrids), supported by population structure analyses (admixture and principal component analysis (PCA)), which also highlighted extensive genetic admixture among hybrids (Fig. 2d,e).

Saccharum spontaneum exhibited higher nucleotide diversity ($\pi = 1.28 \times 10^{-4}$) than *S. officinarum* ($\pi = 2.43 \times 10^{-5}$), reflecting its broader natural distribution (Fig. 2f), and hybrid sugarcane showed diversity ($\pi = 1.27 \times 10^{-4}$) similar to its *S. spontaneum* progenitor, probably due to diverse wild germplasm and gene flow. The population differentiation statistic between *S. spontaneum* and *S. officinarum* ($F_{ST} = 0.20$) is higher than that between *S. spontaneum* and hybrids ($F_{ST} = 0.12$), or between *S. officinarum* and hybrids ($F_{ST} = 0.08$). Moreover, *S. spontaneum* showed a more rapid decay of linkage disequilibrium, probably reflecting its larger effective population size (Fig. 2g).

POJ2878 shaping modern breeding

To access the genetic contribution of POJ2878 to modern sugarcane cultivars, we analysed 573 geographically diverse samples (327 Chinese, 246 non-Chinese). IBD analysis revealed that 95.3% cultivars shared over 821.6 Mb of IBD sequences with POJ2878, underscoring the extensive use in global sugarcane breeding programmes (Fig. 3a). Genome-wide analysis showed that 98.15% of the genome exhibited introgression from POJ2878 (Fig. 3b), with the top 5% of IBD-dense regions spanning 474.70 Mb. These regions contained 28,308 protein-coding allelic or homoeologous genes, representing haplotypes favoured by breeders. These genes were significantly enriched in important biological processes, including starch and sucrose metabolism and environmental adaptation (Fisher's exact test, $P < 0.05$; Supplementary Fig. 16). Key genes included *sucrose synthase* (*SUS*), *sucrose-phosphate phosphatase* (*SPP*), *sucrose phosphate synthase* (*SPS*)²¹ and sucrose transporters (*SWEET14*, *SWEET15* and *SUT*) (Fig. 3c). Allele-specific gene expression analysis further revealed that many of these favourable haplotypes exhibited expression in stem and/or leaf tissues (Fig. 3d). However, their expression was not always the highest or lowest among the alleles, consistent with previous observation²². This pattern suggests that alternative haplotypes may still hold untapped potential for future improvement.

The *SUS2* gene has four haplotypes: one favourable haplotype from the *S. spontaneum* subgenome (*Ss_hap1*) and three from the *S. officinarum*-derived subgenome (*So_hap1–3*). *Ss_hap1* includes a 27-bp insertion in exon 9 and a 255-bp deletion in exon 11 (Fig. 3e). We identified six unique k -mers to this haplotype and used their average copy number to estimate its prevalence in each cultivar. A strong correlation was found between *Ss_hap1* copy number and sucrose content at 270 days after planting, with cultivars carrying 1–3 copies showing substantially higher sucrose levels (Fig. 3f and Supplementary Table 12). However, sucrose content did not increase proportionally with more copies, as excessive *S. spontaneum* subgenome contributions may be detrimental.

Further analysis revealed substantial differences in the use of POJ2878 between Chinese and non-Chinese cultivars. On average, non-Chinese cultivars exhibited a higher IBD density (0.20) compared with Chinese cultivars (0.18; $P < 0.001$, Mann–Whitney U -test; Fig. 3b), indicating broader adoption of POJ2878 as a genetic donor in non-Chinese breeding programmes. Moreover, IBD analysis revealed distinct genomic regions preferentially selected by Chinese and non-Chinese breeders (Fig. 3g), with the top 5% of IBD windows containing 5,057 protein-coding genes in Chinese and 2,838 in non-Chinese cultivars. Functional enrichment analysis revealed divergent selection pressures: genes in Chinese cultivars were significantly enriched for herbicide response traits, whereas those in non-Chinese cultivars were enriched for abiotic stress response (Fisher's exact test, $P < 0.05$; Supplementary Fig. 17). These findings highlight the critical role of

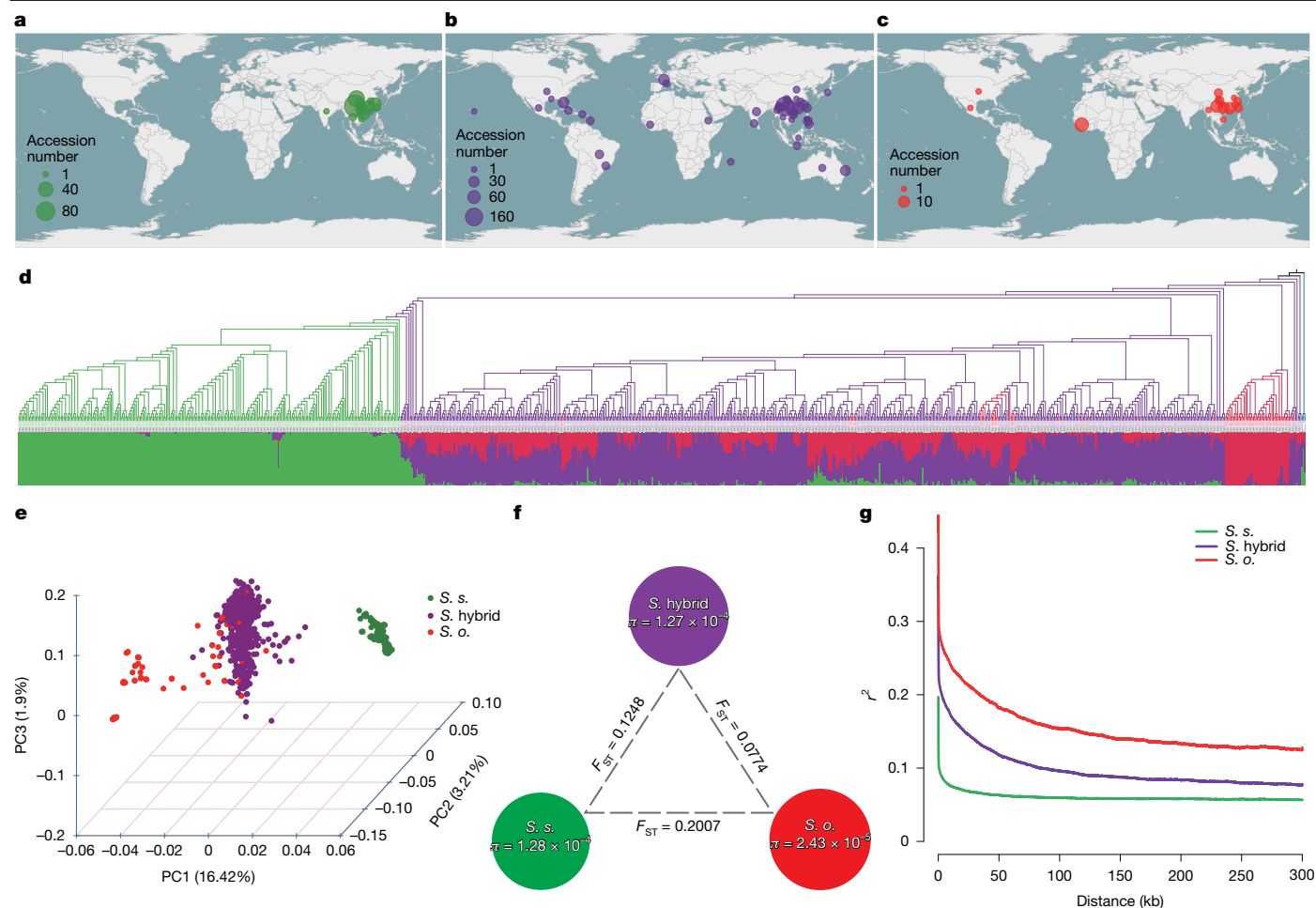


Fig. 2 | Genetic diversity and population structure analysis. a–c, The geographical distribution of *S. spontaneum* (*S. s.*) (**a**; $n = 290$), hybrid cultivars (*S. hybrid*) (**b**; $n = 613$) and *S. officinarum* (*S. o.*) (**c**; $n = 78$), showcasing the global representation of the sampled accessions. The base map was generated using the R package *rnatuarearth* with data from Natural Earth (<http://natuarearthdata.com>). **d**, Phylogenetic tree and admixture analysis ($k = 3$) of the resequenced accessions, illustrating the genetic relationships and levels of admixture among *S. spontaneum*, *S. officinarum* and hybrid populations. **e**, PCA of *S. spontaneum*, *S. officinarum* and hybrid accessions, highlighting

the genetic differentiation and clustering of the three populations. **f**, The genome-wide mean nucleotide diversity (π) within each population and population fixation statistics (F_{ST}) between different populations, revealing the genetic diversity and differentiation among *S. spontaneum*, *S. officinarum* and hybrids. **g**, Linkage disequilibrium decay patterns across the three populations; linkage disequilibrium was measured as r^2 and plotted against physical distance (kb), showing the population-specific difference in linkage disequilibrium decay rates.

POJ2878 in shaping the genetic architecture of modern sugarcane cultivars and underscore region-specific selection pressures in global breeding programmes.

Sugarcane domestication and improvement

To investigate the development of key traits in sugarcane, we analysed selective sweeps during three critical phases: natural selection in *S. spontaneum*, early domestication in *S. officinarum* and crop improvement in hybrid cultivars through nobilization. Natural selection in *S. spontaneum* was examined using within-population comparisons with the *S. spontaneum* 82-114 T2T monoploid assembly as a reference. Early domestication in *S. officinarum* involved comparisons with its direct ancestor, *Saccharum robustum*, using the *S. officinarum* XZ T2T monoploid assembly as a reference. Moreover, the detection of selective sweeps during the crop improvement stage was based on the POJ2878 haplotype-resolved genome assembly.

In *S. spontaneum*, we identified 17.56 Mb selective sweeps, or 2.29% of its 766.92 Mb genome, encompassing 1,600 protein-coding genes (Fig. 4a). These genes were enriched in biological processes related

to defence response, probably contributing to the species' robust resistance (Supplementary Fig. 18). Notable genes included a homologue of *Arabidopsis thaliana* GLP (*AtGLP*), essential for root growth and development, and stress-response genes such as homologues of *AtRPM1* (pathogen resistance)²³ and *AtCBL1* (regulator of drought, salt and cold stress responses²⁴). Overexpression of *Saccharum* hybrid *CBL1* (*ShCBL1*) in sugarcane increased proline content, superoxide dismutase (SOD) activity and malondialdehyde levels ($P < 0.05$, one-way analysis of variance (ANOVA)), and its overexpression in rice affected *COLD1* and *ICE1* (Extended Data Fig. 5), enhancing chilling tolerance^{25,26}. These findings align with the long-standing view that *S. spontaneum* contributed strong ratooning ability and resistance traits to modern sugarcane cultivars.

The domestication of *S. officinarum*, originating in New Guinea from *S. robustum*^{5,27}, involved substantial genetic changes. Comparative selective sweep analyses between *S. officinarum* and *S. robustum* identified 20.22 Mb of domesticated regions, representing 2.30% of the 879.72 Mb genome and encompassing 1,099 protein-coding genes (Fig. 4b). These genes were enriched in biological processes critical to key agronomic traits, particularly carbohydrate metabolic

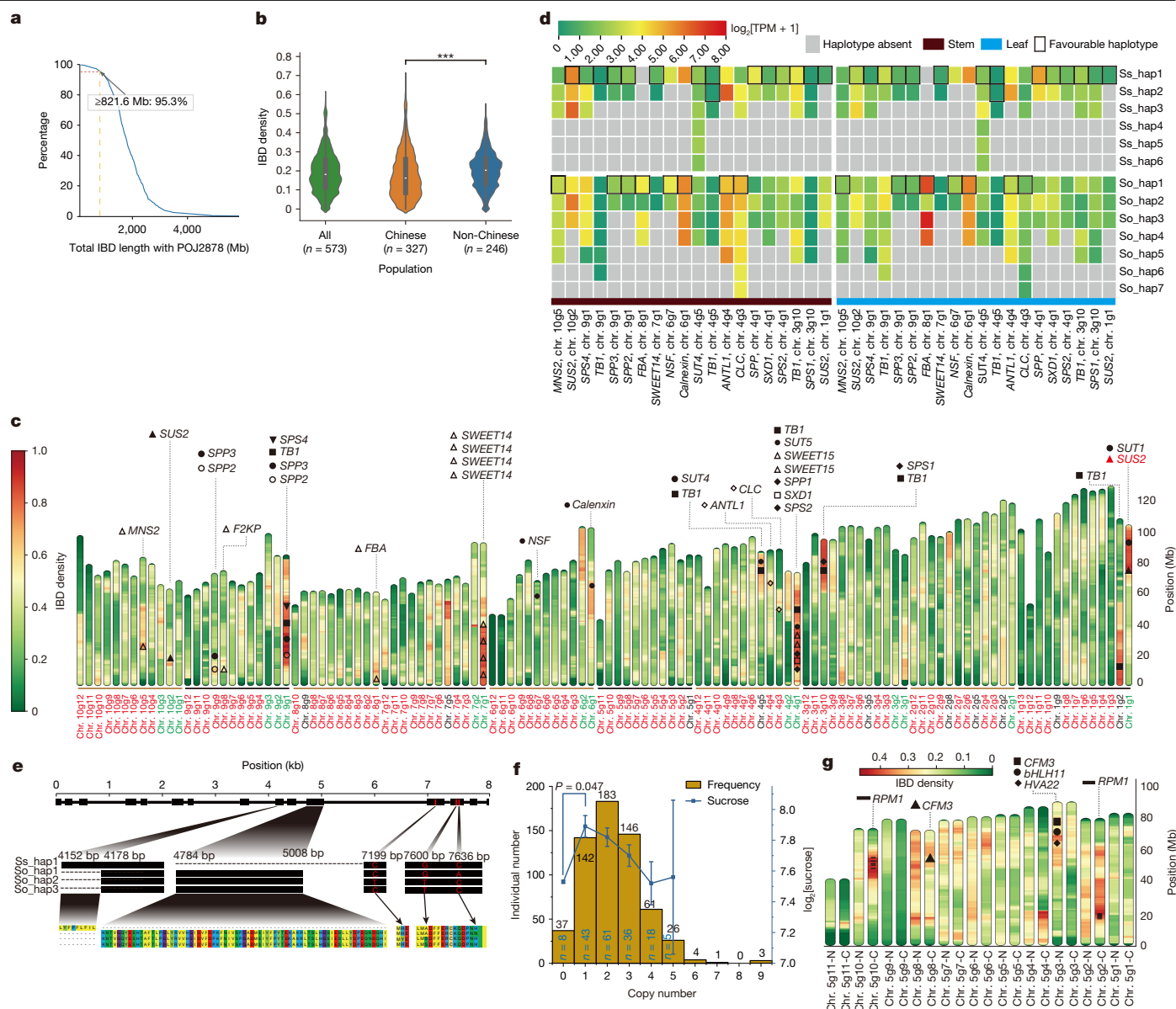


Fig. 3 | The genetic contribution of POJ2878 and analysis of breeder-favoured haplotypes. **a**, Empirical cumulative distribution of total IBD length shared between 573 sugarcane cultivars and POJ2878. The x-axis shows the total IBD length per cultivar (Mb), and the y-axis shows the percentage of cultivars. The dashed line at 821.6 Mb indicates that 95.3% of cultivars share at least 821.6 Mb of IBD sequence with POJ2878. **b**, The IBD density in 50-kb non-overlapping genomic windows across global ($n = 573$), Chinese ($n = 327$) and non-Chinese ($n = 246$) population cultivars. Statistical analysis was performed using two-sided Mann-Whitney U -tests; $***P < 0.001$. The violin plot width reflects the frequency of windows; the white dots show the average IBD density; the box indicates the interquartile range (IQR; 25th–75th percentiles); and the whiskers extend to $1.5 \times \text{IQR}$. Exact P values are provided as Source data. **c**, The distribution of IBD density across the 118 chromosomes. The red regions indicate higher IBD density (that is, breeder-favoured haplotypes), and the green regions indicate lower values. *Saccharum spontaneum*-derived,

and biosynthetic processes (Fisher's exact test, $P < 0.05$; Supplementary Fig. 19). Notably, selected genes included regulators of flowering time (such as *RFT1*²⁸ and *TFL4*²⁹), probably contributing to flowering suppression—a trait that enhances biomass and sugar accumulation. This aligns with findings that homologues of *flowering locus T* in sugarcane can delay flowering time and modify productive structure³⁰. Key genes involved in sugar metabolism and transport, such as *ShFPB*

S. officinarum-derived and recombinant chromosomes are coloured green, red and black, respectively. *SUS2* is highlighted in red to facilitate detailed analysis. **d**, Allele-specific expression shows genes (x axis) with chromosome-specific haplotypes (for example, chromosome 10g5) and progenitor origins (y axis = 3 replicates) from mature stem and leaf were normalized as $\log_2[\text{transcripts per million (TPM)} + 1]$. The black squares highlight breeder-selected haplotypes and the grey areas indicate haplotype absence. **e**, Haplotypic variations of *SUS2*, with *Ss_hap1* identified as the breeder-favoured haplotype. **f**, The frequency of *SUS2Ss_hap1* CNVs across 595 samples; the bars represent sample counts and proportion. The broken line represents the average sucrose content, and the error bars indicate the s.e.m. Significance was assessed using one-way ANOVA. **g**, Differential use of POJ2878 haplotypes exists between Chinese (-C) and non-Chinese (-N) populations.

(fructose-1-6-biophosphatase)³¹ and *Sh/VV* (responsible for breaking down sucrose into glucose and fructose)³², were also under selection, supporting the early domestication of sugarcane.

Selective sweep analyses during the crop improvement process of hybrid cultivars, compared with their progenitor species, identified 4,032 genes (Fig. 4c,d) associated with response to light, carbohydrate metabolic process, response to stimulus and shoot system

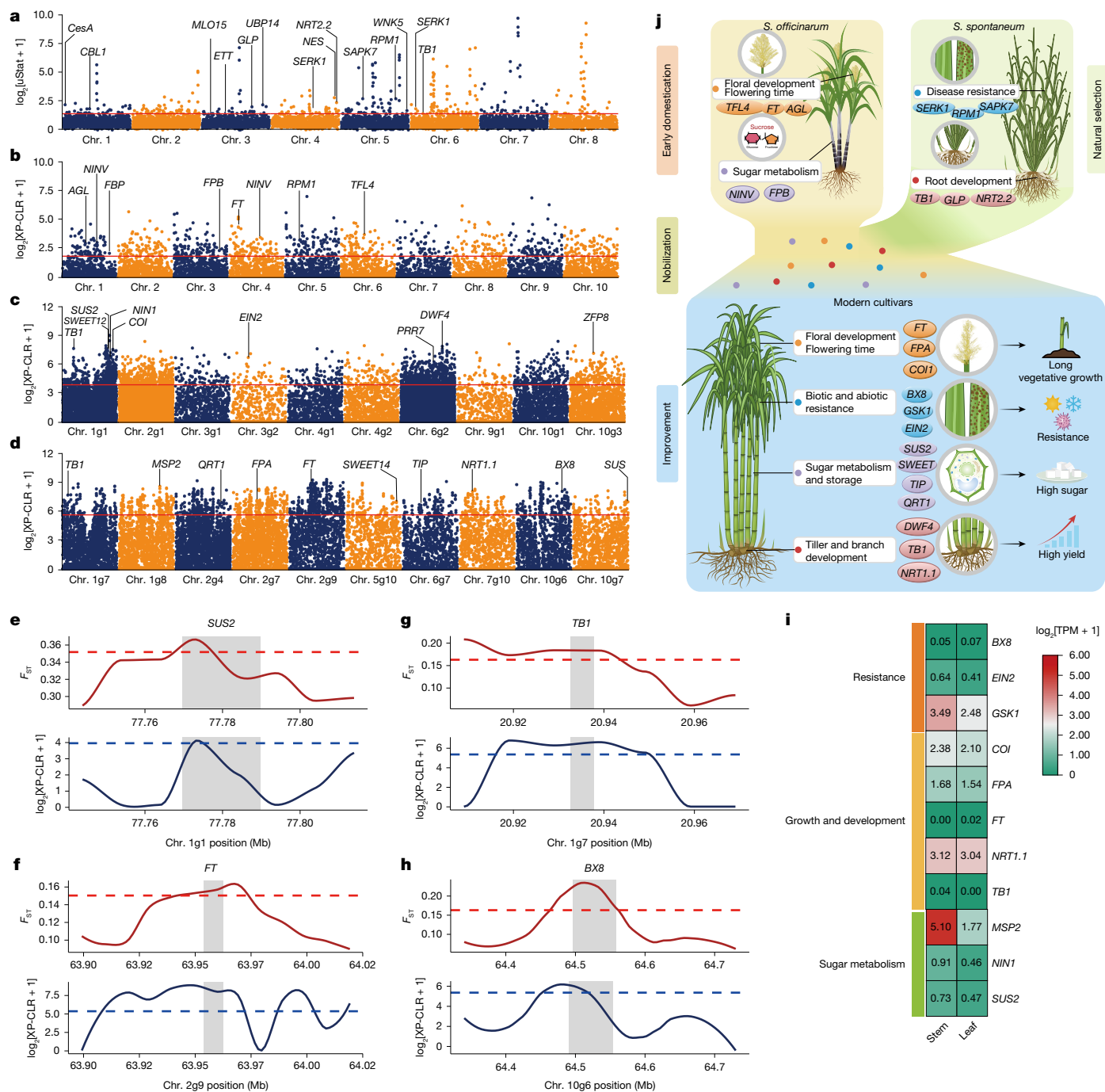


Fig. 4 | Analysis of selective sweeps in *S. spontaneum*, *S. officinarum* and *S. hybrid*. **a**, Genome-wide distribution of selective sweep signals identified through within-population comparisons in the *S. spontaneum* population ($n = 290$ independent varieties). The T2T monoploid genome of *S. spontaneum* 82-114 was used as the reference genome. **b**, Selective sweep signals identified through cross-population comparisons in the *S. officinarum* population ($n = 78$ independent varieties) against the *S. robustum* population ($n = 43$ independent varieties). The T2T monoploid genome of *S. officinarum* XZ was used as the reference genome. **c, d**, Selective sweep signals identified through cross-population comparisons with *S. spontaneum* (**c**; $n = 290$) and *S. officinarum* (**d**; $n = 78$) in the *S. hybrid* population ($n = 613$). Sweep candidates were defined as genomic windows in the top 5% of genome-wide uStat (**a**) or XP-CLR (**b–d**) values. Genes with putative functional relevance are highlighted in the Manhattan plots. **e–h**, Local population fixation statistics (F_{ST}) and

XP-CLR profiles at selected target loci (*SUS2* (**e**), *FT* (**f**), *TB1* (**g**) and *BX8* (**h**)), highlighting regions under strong selection in the *S. hybrid* population. **i**, Expression analysis of allelic genes within selective swept regions related to growth development, vacuole development and disease resistance under selective sweeps. Stem and leaf tissues were collected with three biological replicates per tissue ($n = 3$). Expression levels are shown as $\log_2[TPM + 1]$. **j**, A proposed model delineating natural and artificial selection in sugarcane. In wild sugarcane (*S. spontaneum*), genes associated with disease resistance and root development are subject to natural selection. During the early domestication of *S. officinarum*, selective sweeps targeted genes linked to flowering time and sugar metabolism. After nobilization breeding, crop improvement efforts in modern sugarcane hybrids (*S. hybrid* cultivars) have driven selective sweeps in genes governing floral development, flowering time, biotic and abiotic resistance, vacuole development and tiller architecture.

development (Supplementary Fig. 20). Sugar accumulation and resistance related genes were identified, including homologues of *SUS2* (Fig. 4e), *SWEET14*, *TPS11*³³, *BX8*³⁴ (Fig. 4h) and *GSKI*³⁵. Key genes also include homologues of *TB1* (Fig. 4g), which reduces branching for upright plant structure³⁶, *FT* (flowering locus T; Fig. 4f) and *NRT1.1*, which facilitates lateral root branching by controlling the transport of the hormone auxin in response to nitrate availability³⁷. Notably, genes under strong artificial selection for vacuole development and cell wall loosening include homologues of *TIP1*³⁸ and *QRT1*³⁹. *TIP1* encodes an aquaporin protein that facilitates water transport across the vacuole membrane in *Arabidopsis*³⁸. Overexpression of Sh*TIP1* modified leaf width in sugarcane and rice by reducing vein distance without affecting veins and epidermal cell count (Extended Data Fig. 6), suggesting that it has a role in regulating cell size during development. The *QRT1* gene encodes a pectin methylesterase in *Arabidopsis* that is involved in cell wall loosening and modification³⁹.

To confirm that the selection signals (Fig. 4e–h) exceed the expectations of random drift, we compared the cross-population composite likelihood ratio (XP-CLR) and F_{ST} values of candidate genes against genome-wide empirical distributions derived from all gene regions. Candidate genes (*SUS2*, *FT*, *TB1* and *BX8*) were identified as significant outliers ($P < 0.05$, Fisher's combined test), with their values in the extreme tails of the empirical distributions (Supplementary Fig. 21). Expression analysis supported their potential functions during the natural and/or artificial selection (Fig. 4i). Together, these genes underpin sugarcane domestication and crop improvement, contributing to traits such as flowering regulation, stress response and sucrose metabolism and storage—key characteristics that define modern sugarcane as a leading sugar crop (Fig. 4j).

GWAS reveals key genes for sugar storage

Sugarcane's ability to maintain high sugar content is largely attributed to its efficient sucrose storage in stem parenchyma cells⁴⁰. Our analysis revealed that parenchyma cells in high-sugar-content *S. officinarum* (10,525 μm^2) are significantly larger than those in *S. spontaneum* 82-114 (6,526 μm^2 , Mann–Whitney *U*-test, $P < 0.01$; Fig. 5a–f). Similarly, high-sugar-content hybrids had larger cells compared with low-sugar-content hybrids (14,862 μm^2 versus 6,506 μm^2). To further examine this relationship, we measured parenchyma cell density and size across 265 hybrid cultivars using a robust image recognition approach (Extended Data Fig. 7). Significant correlations were observed between these cell traits and Brix (a measure of sugar content; $P < 0.05$; Supplementary Fig. 22), supporting the potential breeding value of parenchyma cells.

Traditional GWASs face comprehensive limitations in polyploid genomes including inability to detect multiallelic variants, neglecting allele dosage effects and inaccuracies in genotyping due to ambiguous short-read alignments. To overcome these challenges, we used KMERIA¹⁵, a *k*-mer-based GWAS approach specifically designed for polyploid genomes (Methods). KMERIA analysis of parenchyma cell traits and sugar content revealed key genes linked to important biological functions (Fig. 5g–i and Supplementary Table 13). These include a homologue of *AtMob1A* that is essential for organ growth and tissue patterning⁴¹, which may regulate cell proliferation and organ size⁴². We also identified a MATE (multidrug and toxic compound extrusion) transporter family member that is involved in hormone transport and stress responses, highlighting their potential in parenchyma cell development⁴³. Moreover, we identified a WD40-domain-containing protein homologous to the cucumber *LITTLELEAF* (*LL*) gene, which is known to control organ size and lateral branching⁴⁴. This gene may similarly influence parenchyma cell size and contribute to sugarcane's architecture and sucrose storage.

Notably, the *SUT2* (sucrose transporter 2), a gene that is essential for sucrose transport, was identified by KMERIA to have a significant

association with parenchyma cell size (Fig. 5g). Overexpression of Sh*SUT2* in sugarcane cultivars significantly increased parenchyma cell size, with a 7.20% enhancement in the OE1 line and a 28.92% increase in the OE2 line ($P < 0.0001$, Wilcoxon rank-sum tests; Extended Data Fig. 8).

Discussion

This study establishes a comprehensive genomic framework (Supplementary Note 4 and Extended Data Fig. 9) for understanding the genetic basis of sugarcane domestication and improvement. By generating a fully haplotype-phased assembly of the hybrid cultivar POJ2878, which substantially reduced switch errors and collapsed assembly artefacts present in previous assemblies, we identified key genetic innovations that have driven the transformation of sugarcane from wild *Saccharum* species into a high-sugar, high-biomass crop. In *S. officinarum*, genes under domestication involved in regulating sucrose metabolism (*NINV*, *FPB*) and floral development (for example, *TFL4* and *AGL80*), consistent with selection for delayed flowering and enhanced sink strength. By contrast, *S. spontaneum* retained alleles associated with disease resistance and ratooning ability, including homologues of *RPMT*, *SERKT*⁴⁵ and *TB1*, which were later introgressed into hybrids during nobilization. We demonstrate that POJ2878 occupies a central position in the sugarcane improvement trajectory. More than 95% of modern cultivars harbour extensive IBD tracts from POJ2878, and around 98% of the genome shows non-zero introgression across the panel, including haplotypes at *SUS2*, *SWEETs*, *SPSs* and *SUTs*. At the same time, we observe that increasing dosage of the favourable *S. spontaneum*-derived *SUS2* haplotype improves sucrose content only up to a point; excessive *S. spontaneum* contribution, typically reflecting low nobilization, is associated with reduced agronomic performance. This pattern highlights a general principle of polyploid breeding: beneficial haplotypes often act within an optimal dosage window, and overall subgenome balance cannot be ignored. A haplotype-resolved reference and dosage-aware approaches such as KMERIA provide a practical framework for explicitly managing this trade-off.

Our results also connected cellular anatomy to agronomic performance. Differences in parenchyma cell size and density between high- and low-sugar lines, combined with KMERIA signals at *Mob1A*, *MATE* transporters, LL-like WD40 proteins and *SUT2*, indicate that vacuole development and sink cell architecture represent underexplored targets for improving sugar yield. Overexpression of Sh*SUT2* increased parenchyma cell size, supporting a causal relationship between sucrose transport capacity and sink cell expansion. Nonetheless, given the complexity of sugarcane's polyploid genome and possible functional divergence among homoeologues, these inferences should be regarded as testable hypotheses that require systematic validation in sugarcane germplasm and field conditions.

This study provides a template for genomics-assisted breeding in sugarcane and other polyploid crops. In practice, breeders can now (1) track and optimize the dosage of favourable haplotypes such as the *S. spontaneum*-derived *SUS2* allele; (2) select or engineer *Mob1A*, *MATE* or LL-like haplotypes to fine-tune parenchyma cell proliferation and organ architecture; and (3) exploit POJ2878-derived regions while monitoring genetic diversity to avoid over-narrowing the breeding base.

The contrasting selection signatures observed in Chinese versus non-Chinese cultivars, for example, enrichment of genes associated with herbicide response versus abiotic stress tolerance, highlight how regional priorities shape the genomic architecture of crop improvement. These patterns support the need for locally tailored genomic strategies that integrate climatic conditions, management practices and market demands, rather than a one-size-fits-all deployment of elite alleles. At the same time, the widespread global reliance on POJ2878 underscores both the value and the potential vulnerability associated with foundational cultivars in polyploid breeding programmes.

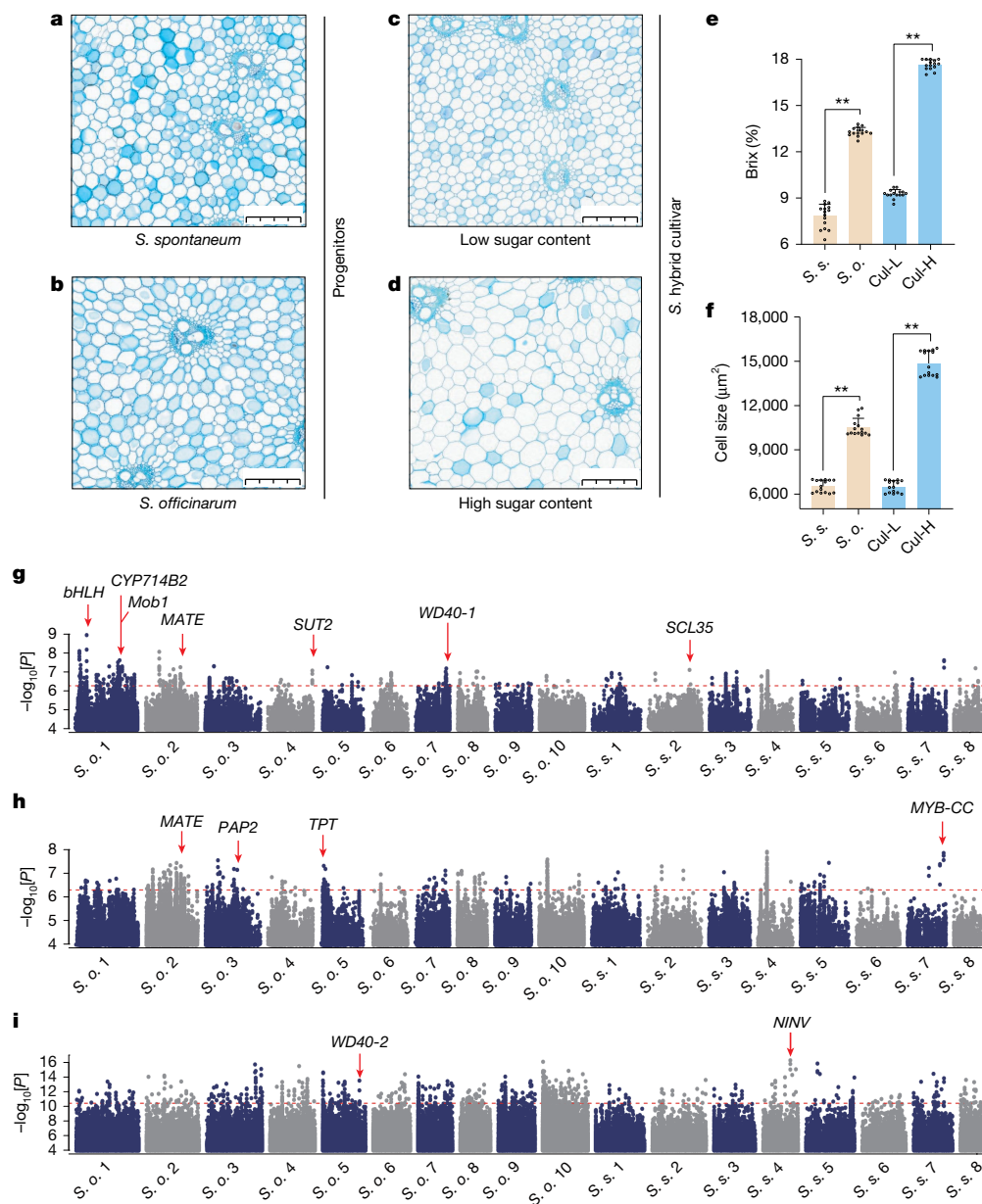


Fig. 5 | Genome-wide association analysis of parenchyma cell traits.

a–d, Representative stem cross-sections showing parenchyma cell organization in four groups: the two progenitor species (*S. spontaneum* (**a**) and *S. officinarum* (**b**)), low-sugar-content *S.* hybrid cultivars (Cul-L) (**c**) and high-sugar-content *S.* hybrid cultivars (Cul-H) (**d**). One representative image is shown for each group. For each group, the microscopy observations were repeated three times independently with similar results. For **a–d**, scale bars, 500 μm . **e**, Brix (%) measurements across the four groups. For each biological sample, Brix was measured 15 times ($n = 15$ independent biological replicates). Data are mean $\pm$ s.d. Statistical significance was assessed using a two-sided Student's *t*-test. **f**, Parenchyma cell size across the four groups quantified from

15 microscopy fields of view per biological sample. $n = 15$ fields of view per group. Data are mean $\pm$ s.d. Statistical significance was assessed using a two-sided Student's *t*-test. **g–i**, Genome-wide association results for parenchyma cell size (μm^2) (**g**), cell density (cells per mm^2) (**h**) and Brix (%) (**i**) in 265 cultivars ($n = 265$). The y axis shows $-\log_{10}[P]$, where the *P* values indicate significance of *k*-mer–trait associations, determined using two-sided Wald tests in a linear mixed model using KMERIA. The horizontal dashed lines indicate the Benjamini–Hochberg false-discovery-rate-corrected significance threshold with *P* values of 5.10×10^{-7} (**g**), 5.06×10^{-7} (**h**) and 3.85×10^{-11} (**i**), respectively. Exact *P* values for **e** and **f** are provided as Source data. ***P* < 0.01.

More broadly, resolving the genomic architecture of sugarcane bridges a critical gap between polyploid genome structure and agronomic function. By integrating long-read and proximity-ligation sequencing, haplotype-aware assembly, *k*-mer-based association and allele-specific expression profiling, this study establishes a generalizable framework for dissecting complex traits in polyploid crops. The genes, haplotypes and pathways identified here provide immediate targets for improving sugar yield and resilience in sugarcane. As demands for sustainable sugar and bioenergy production intensifies,

such polyploid-aware genomic resources and analytical tools will be indispensable for accelerating the development of climate-resilient, high-efficiency cultivars in sugarcane and other staple crops.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions

1. D'Hont, A. & Glaszmann, J. C. Sugarcane genome analysis with molecular markers, a first decade of research. *Proc. Int. Soc. Sugarcane Technol.* **24**, 556–559 (2001).
2. Li, N. et al. BvBZRI improves parenchyma cell development and sucrose accumulation in sugar beet (*Beta vulgaris* L.) taproot. *Front. Plant Sci.* **16**, 1495161 (2025).
3. Wang, K. et al. Overexpression of tonoplast transporter FvMATE51 simultaneously increases fruit size and sugar accumulation in strawberry. *Plant Biotechnol. J.* **24**, 2538–2540 (2025).
4. Moore, P. H. & Cosgrove, D. J. Developmental changes in cell and tissue water relations parameters in storage parenchyma of sugarcane. *Plant Physiol.* **96**, 794–801 (1991).
5. Garsmeur, O. et al. The genomic footprints of wild *Saccharum* species trace domestication, diversification, and modern breeding of sugarcane. *Cell* **188**, 7252–7266 (2025).
6. Bao, Y. et al. A chromosomal-scale genome assembly of modern cultivated hybrid sugarcane provides insights into origination and evolution. *Nat. Commun.* **15**, 3041 (2024).
7. Healey, A. L. et al. The complex polyploid genome architecture of sugarcane. *Nature* **628**, 804–810 (2024).
8. Zhang, J. et al. The highly allo-autopolyploid modern sugarcane genome and very recent allopolyploidization in *Saccharum*. *Nat. Genet.* **57**, 242–253 (2025).
9. Huang, Y. et al. Multiscale pangenome graphs empower the genomic dissection of mixed-ploidy sugarcane species. *Science* **391**, eadx1616 (2026).
10. Kong, W., Wang, Y., Zhang, S., Yu, J. & Zhang, X. Recent advances in assembly of complex plant genomes. *Genom. Proteom. Bioinform.* **21**, 427–439 (2023).
11. Amarasinghe, S. L. et al. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 30 (2020).
12. Deshpande, A. S. et al. Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. *Nat. Biotechnol.* **40**, 1488–1499 (2022).
13. Al'Khafaji, A. M. et al. High-throughput RNA isoform sequencing using programmed cDNA concatenation. *Nat. Biotechnol.* **42**, 582–586 (2024).
14. Wang, Y. et al. Enhanced Pore-C with C-Phasing enables chromosomal-scale, haplotype-resolved assembly of ultra-complex genomes. Preprint at *Research Square* <https://doi.org/10.21203/rs.3.rs-7343323/v1> (2025).
15. Chen, S. et al. A k-mer-based GWAS approach empowering gene mining in polyploids. Preprint at *Research Square* <https://doi.org/10.21203/rs.3.rs-7347406/v1> (2025).
16. Zhang, J. et al. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nat. Genet.* **50**, 1565–1573 (2018).
17. Adamowski, M. & Friml, J. PIN-dependent auxin transport: action, regulation, and evolution. *Plant Cell* **27**, 20–32 (2015).
18. Zhang, H. et al. *Carbon Starved Anther* encodes a MYB domain protein that regulates sugar partitioning required for rice pollen development. *Plant Cell* **22**, 672–689 (2010).
19. Maurel, C., Verdoucq, L., Luu, D.-T. & Santoni, V. Plant aquaporins: membrane channels with multiple integrated functions. *Annu. Rev. Plant Biol.* **59**, 595–624 (2008).
20. Wang, W., Kim, J., Martinez, T. S., Huq, E. & Sung, S. COP1 controls light-dependent chromatin remodeling. *Proc. Natl Acad. Sci. USA* **121**, e2312853121 (2024).
21. Anur, R. M., Mufithah, N., Sawitri, W. D., Sakakibara, H. & Sugiharto, B. Overexpression of sucrose phosphate synthase enhanced sucrose content and biomass production in transgenic sugarcane. *Plants* **9**, 200 (2020).
22. Kremling, K. A. G. et al. Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature* **555**, 520–523 (2018).
23. Grant, M. et al. The *RPM1* plant disease resistance gene facilitates a rapid and sustained increase in cytosolic calcium that is necessary for the oxidative burst and hypersensitive cell death. *Plant J.* **23**, 441–450 (2000).
24. Cheong, Y. H. et al. CBL1, a calcium sensor that differentially regulates salt, drought, and cold responses in *Arabidopsis*. *Plant Cell* **15**, 1833–1845 (2003).
25. Ma, Y. et al. *COLD1* confers chilling tolerance in rice. *Cell* **160**, 1209–1221 (2015).
26. Chinnusamy, V. et al. ICE1: a regulator of cold-induced transcriptome and freezing tolerance in *Arabidopsis*. *Genes Dev.* **17**, 1043–1054 (2003).
27. Okemo, P. et al. Crop domestication in the Asia Pacific Region: a review. *Agric. Commun.* **2**, 100032 (2024).
28. Hua, H. et al. A novel non-functional *RFT1* allele delays heading date in rice. *Rice* **18**, 106 (2025).
29. Zhu, Y. et al. TERMINAL FLOWER 1-FD complex target genes and competition with FLOWERING LOCUS T. *Nat. Commun.* **11**, 5118 (2020).
30. Coelho, C. P., Minow, M. A. A., Chalfun-Jão, A. & Colasanti, J. Putative sugarcane *FT/FTL1* genes delay flowering time and alter reproductive architecture in *Arabidopsis*. *Front. Plant Sci.* **5**, 221 (2014).
31. Groenewald, J.-H. & Botha, F. C. Molecular and kinetic characterisation of sugarcane pyrophosphate: fructose-6-phosphate 1-phosphotransferase and its possible role in the sucrose accumulation phenotype. *Funct. Plant Biol.* **34**, 517–525 (2007).
32. Lai, R. et al. Identification of the sugarcane invertase gene family with deciphering the key role of *ShN/AINV3.1* in drought stress response. *Plant Stress* **20**, 101261 (2026).
33. Singh, V., Louis, J., Ayre, B. G., Reese, J. C. & Shah, J. *TREHALOSE PHOSPHATE SYNTHASE11*-dependent trehalose metabolism promotes *Arabidopsis thaliana* defense against the phloem-feeding insect *Myzus persicae*. *Plant J.* **67**, 94–104 (2011).
34. Li, T. et al. From defense chemistry to agroecological solutions: the evolving role of benzoxazinoids in plants. *Crop J.* **13**, 1333–1349 (2025).
35. Piao, H. L., Lim, J. H., Kim, S. J., Cheong, G. & Hwang, I. Constitutive overexpression of AtGSK1 induces NaCl stress responses in the absence of NaCl stress and results in enhanced NaCl tolerance in *Arabidopsis*. *Plant J.* **27**, 305–314 (2001).
36. Pribil, M. et al. Altering sugarcane shoot architecture through genetic engineering: prospects for increasing cane and sugar yield. *Proc. Aust. Soc. Sugar Cane Technol.* **29**, 251–257 (2007).
37. Krouk, G. et al. Nitrate-regulated auxin transport by NRT1.1 defines a mechanism for nutrient sensing in plants. *Dev. Cell* **18**, 927–937 (2010).
38. Jauh, G.-Y., Fischer, A. M., Grimes, H. D., Ryan, C. A. & Rogers, J. C. Tonoplast intrinsic protein defines unique plant vacuole functions. *Proc. Natl Acad. Sci. USA* **95**, 12995–12999 (1998).
39. Francis, K. E., Lam, S. Y. & Copenhaver, G. P. Separation of *Arabidopsis* pollen tetrads is regulated by *QUARTET1*, a pectin methylesterase gene. *Plant Physiol.* **142**, 1004–1013 (2006).
40. Wang, J., Nayak, S., Koch, K. & Ming, R. Carbon partitioning in sugarcane (*Saccharum* species). *Front. Plant Sci.* **4**, 201 (2013).
41. Pinoso, F. et al. The *Arabidopsis thaliana* *Mob1A* gene is required for organ growth and correct tissue patterning of the root tip. *Ann. Bot.* **112**, 1803–1814 (2013).
42. Xiong, J. et al. The Hippo/STE20 homolog SIK1 interacts with MOB1 to regulate cell proliferation and cell expansion in *Arabidopsis*. *J. Exp. Bot.* **67**, 1461–1475 (2016).
43. Gani, U., Vishwakarma, R. A. & Misra, P. Membrane transporters: the key drivers of transport of secondary metabolites in plants. *Plant Cell Rep.* **40**, 1–18 (2021).
44. Yang, L. et al. *LITTLELEAF* (LL) encodes a WD40 repeat domain-containing protein associated with organ size variation in cucumber. *Plant J.* **95**, 834–847 (2018).
45. Hu, H., Xiong, L. & Yang, Y. Rice *SERK1* gene positively regulates somatic embryogenesis of cultured cell and host defense response against fungal infection. *Planta* **222**, 107–117 (2005).

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



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026

¹State Key Laboratory of Tropical Crop Breeding, Key Laboratory of Biology and Genetic Resources of Tropical Crops Ministry of Agriculture and Rural Affairs, Key Laboratory of Conservation and Utilization of Tropical Agricultural Biological Resources of Hainan Province, Institute of Tropical Bioscience and Biotechnology, Chinese Academy of Tropical Agricultural Sciences, Haikou, China. ²State Key Laboratory for Tropical Crop Breeding, Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China. ³State Key Laboratory for Tropical Crop Breeding, Sugarcane Research Institute, Yunnan Academy of Agricultural Sciences, Kaiyuan, China. ⁴Center for Genomics and Biotechnology, National Engineering Research Center for Sugarcane, Fujian Agriculture and Forestry University, Fuzhou, China. ⁵Guangxi Key Laboratory for Sugarcane Biology, Guangxi University, Nanning, China. ⁶South Subtropical Crop Research Institute, Chinese Academy of Tropical Agricultural Sciences, Zhanjiang, China. ⁷Sugarcane Research Institute, Guangxi Academy of Agricultural Sciences, Nanning, China. ⁸Institute of Nanfan and Seed Industry, Guangdong Academy of Sciences, Guangzhou, China. ⁹Key Laboratory for Biosafety Monitoring and Molecular Breeding in Off-Season Reproduction Regions of Hainan Province, Sanya Research Institute of Chinese Academy of Tropical Agricultural Sciences, Sanya, China. ¹⁰College of Geography and Oceanography, Minjiang University, Fuzhou, China. ¹¹These authors contributed equally: Jungang Wang, Xiaofeng Li, Yibin Wang, Jishan Lin, Shuai Chen, Hongbo Liu, Xiao Chen. ✉e-mail: zmuqing@163.com; rayming@illinois.edu; zhangxingtan@caas.cn

Article

Methods

Plant materials

For genome assembly and annotation, we used plant materials including, *S. hybrid* POJ2878, *S. spontaneum* 82-114 and *S. officinarum* XZ. These were cultivated in the greenhouse at Shenzhen. Tender leaves were collected for subsequent genome sequencing.

This study also resequenced 981 sugarcane accessions, including 78 *S. officinarum*, 290 *S. spontaneum* accessions and 613 cultivated varieties. Among the cultivated varieties, 246 were non-Chinese germplasms, while 327 were core domestic cultivars or breeding lines in China. The plant materials were grown at three experimental sites representing distinct climates: (1) a tropical monsoon climate at the sugarcane research base in Lingao County, Hainan Province (19° 34′–20° 02′ N, 109° 3′–109° 53′ E, altitude: 34 m); (2) a subtropical monsoon climate in Fusui County, Guangxi Zhuang Autonomous Region (22° 38′ 06″ N, 107° 54′ 15″ E; altitude 94 m); and (3) a subtropical plateau climate in Kaiyuan City, Yunnan Province (23° 30′–23° 58′ N, 103° 04′–103° 43′ E; altitude 1,051 m). Buds were first disinfected and then planted using a single-bud method within a randomized complete block design (1.2 m row spacing; 15 cm plant spacing). To ensure uniform germination, the field was subject to standardized management, including regular irrigation and pest and weed control.

DNA and RNA sequencing

DNA extraction. DNA extraction for single-molecule sequencing involved collecting samples and preparing long genomic DNA. The DNA was then purified using the QIAGEN Genomic Kit. For ONT-UL sequencing, the SDS method was used to extract DNA, omitting the purification step to preserve long DNA molecules. The DNA integrity and purity were evaluated using the NanoDrop system. Moreover, the Qubit 4.0 Fluorometer was used to measure DNA concentration.

PacBio sequencing. The library preparation process included gDNA shearing, DNA damage repair, end repair, A-tailing, hairpin adapter ligation, nuclease treatment, size selection and polymerase binding. A total of 15 µg of genomic DNA was used for each sample and sheared using g-TUBEs. We removed single-strand overhangs and performed repair, end-repair and A-tailing on the DNA fragments. Subsequently, we ligated the fragments with hairpin adapters, treated them with nuclease and purified the samples. We used the BluePippin system to select fragments with target sizes and the library size was verified in Agilent 2100. HiFi reads were sequenced on the PacBio Sequel Revio platform.

ONT-UL sequencing. Genomic DNA (8–10 µg per sample) was selected for fragments larger than 50 kb using the SageHLS HMW library system. The ONT-UL reads were sequenced on the PromethION platform.

RNA-seq. High-quality RNA (RIN > 8), extracted and purified using the Omega Bio-tek and Takara kits, was used to construct libraries (NEBNext Ultra) for NovaSeq paired-end sequencing after Agilent 2100 validation.

Isoform sequencing. The full-length RNA libraries were constructed based on the Iso-Seq Express 2.0 kit and the Kinnex kit. The cDNA synthesis was performed, followed by PCR amplification with barcoded Iso-Seq primers. The barcoded full-length cDNA products were pooled and sequenced on the PacBio Revio platform.

Pore-C library construction and sequencing. The Pore-C library was constructed using an enhanced Pore-C (ePore-C) protocol developed in our companion study¹⁴. The ePore-C protocol largely follows the previously published Pore-C methodology¹², with several refinements designed to improve data yield. In brief, approximately 2 g of young plant tissue was finely ground in liquid nitrogen and mixed with a prechilled extraction buffer. Cross-linking was initiated by adding

formaldehyde to the homogenate, incubating for 25 min with periodic vacuum release and then stopped with glycine. The mixture was filtered and the filtrate was layered onto a sucrose buffer and centrifuged to isolate nuclei, followed by further purification using gradient centrifugation with sucrose and Percoll. The nuclei were washed with PBS.

Cross-links were reversed using proteinase K, SDS, Tween-20 and RNase A at 56 °C for 18 h. The sample was divided for DNA extraction using phenol–chloroform, and the aqueous phase was combined and precipitated with NaCl, sodium acetate and ethanol. The DNA pellet was washed, air-dried and resuspended in water. Fragments larger than 1 kb were selected using DNA Clean Beads. DNA concentration and quality were assessed with Qubit and confirmed by agarose gel electrophoresis.

Using a modified Nanopore Ligation Sequencing Kit (2.5 µg input; around 1 µg yield), DNA underwent repair and end-prep at 20 °C and 65 °C and ligation and clean-up at 30 °C before flow cell loading for sugarcane genome analysis.

Genome assemblies

Contig-level assembly. Approximately 30× HiFi reads, combined with ONT-UL reads exceeding 100 kb, were processed using the hifiasm⁴⁶ (v.0.19.1-r559) program using its default settings. The resulting `p_utg` file, which contained haplotype-phased contigs, was then used for subsequent haplotype-resolved scaffolding.

Chromosomal-scale and haplotype-resolved assembly using C-Phasing. The complex hybrid sugarcane genome assembly was successfully achieved using a haplotype-resolved genome scaffolding program, C-Phasing, which was developed in our companion study¹⁴. This program comprises five core modules: Hitig, Methalign, Hyperpartition, Kprune, and Scaffolding. In brief, Hitig uses split-read alignment with ONT Ultra-long reads to identify and correct chimeric and collapsed errors, which are critical for accurate haplotype phasing. Methalign leverages the long reads of Pore-C data to improve alignment accuracy by combining traditional sequence alignment with allele-specific 5mC site information to filter ambiguous mappings and enhance read alignment accuracy. Kprune identifies allelic contigs through pairwise alignments and uses a maximum-matching algorithm to detect and remove interallelic and crossallelic signals, retaining only intrahaplotype interactions for further partitioning. Hyperpartition implements a two-layer partitioning strategy using the Louvain community detection algorithm⁴⁷. It processes the raw hypergraph from Methalign to group allelic contigs into homologous chromosome groups (HCGs). By reweighting edges between allelic and cross-allelic contigs, it refines the hypergraph for accurate haplotype phasing. Scaffolding integrates functionalities from HapHiC⁴⁸ and ALLHiC⁴⁹ to efficiently sort and assemble contig groups using parallel computing. This module addresses the challenge of missing signals in homologous regions, ensuring consistent chromosome orientation across haplotypes. The combined approach substantially enhanced the phasing results, which were then manually corrected using Pore-C interactions in Juicebox⁵⁰. This process ultimately yielded a high-quality, haplotype-resolved genome assembly for hybrid sugarcane.

Assessment of genome assembly

We compared the quality of the POJ2878 genome assembly with three published hybrid sugarcane cultivars (R570 (ref. 7), ZZ1 (ref. 6), and XTT22 (ref. 8)) based on three key criteria: continuity, completeness and accuracy.

Continuity. We measured genome continuity by calculating the contig N50 and the number of gaps for each assembly. This enabled us to compare the structural continuity of the four sugarcane genome assemblies.

Completeness. To evaluate genome completeness, we first used the BUSCO⁵¹ tool with the embryophyta_odb10 database to assess the presence of conserved single-copy orthologous genes. Moreover, we applied Meryl⁵² (v.1.4.1) to generate *k*-mer databases from next-generation sequencing reads by splitting them into 21-mers. Using these *k*-mer databases, we calculated *k*-mer completeness using Merqury⁵² (v.1.3). Using LTR_FINDER_parallel⁵³ (v.1.2) and LTR_HARVEST_parallel (v.1.1; https://github.com/oushujun/LTR_HARVEST_parallel), we identified LTR elements in each genome, which were further processed using LTR_retriever⁵⁴ (v.2.9.6). This library was used to annotate LTRs across the genome and calculate the proportion of intact LTRs relative to total LTRs. To provide a genome-wide measure of LTR completeness, we used the raw LTR Assembly Index⁵⁵. Furthermore, we evaluated genome completeness by analysing the alignment rate of next-generation sequencing reads to the assemblies, the scaffold anchoring rate and the ratio of assembled genome size to estimated genome size. These metrics collectively provided a comprehensive view of the completeness of the four sugarcane genome assemblies.

Accuracy. The base-level accuracy was calculated by Merqury (v1.3). We then assessed structural accuracy using three metrics: Switch error, Collapse error and Hi-C H-trans error.

Switch errors: we extracted the 1,000 longest ultra-long ONT reads and divided them into 5-kb windows. Each window was searched back to the genome using blastn⁵⁶ (v.2.9.0), and windows with a primary alignment score at least 1.5× higher than the secondary alignment (and exceeding a score of 1,000) were considered to be valid. Reads with at least five valid windows were retained for analysis. If at least 50% of the valid windows aligned to a single chromosome, the read was assigned to that chromosome. Any valid window that failed to align to the same chromosome was classified as a switch error. The switch error rate was defined as the proportion of switch error windows relative to all valid windows.

Collapsed regions: to identify collapsed regions, we divided the genome into 50-kb non-overlapping windows. Mosdepth⁵⁷ (v.0.3.9) was used to calculate the depth of each window, denoted as x_i . The set of all x_i values is represented as X , and the frequency of each x_i is denoted as $P(x_i)$. The mode of X was then determined based on the following formula:

$$\text{Mode}(X) = \{x_i | P(x_i) = \max(P(x_j)), \forall x_j \in X\}$$

If x_i exceeds 1.5× mode(X), the depth of that window is considered abnormally high due to collapse. The extent of genomic collapse was quantified by dividing the total size of windows exhibiting abnormal depths by the size of the entire genome. This ratio is defined as the collapse error.

Hi-C H-trans error: Hi-C alignments by BWA⁵⁸ with a mapping quality score (MAPQ) of 0 were discarded to ensure reliable mapping. We then used C-Phasing to convert the BAM files into contact maps, capturing interaction frequencies between genomic regions. For each pair of homologous chromosomes i and j , we defined intrachromosomal interaction counts as C_{ii} and C_{jj} , and interchromosomal interaction counts as C_{ij} . The Hi-C H-trans error for the chromosome pair was calculated using the formula:

$$E_{ij} = \frac{C_{ij}}{(C_{ii} + C_{ij} + C_{jj})}$$

This error rate reflects the proportion of interactions that erroneously link different haplotypes or non-homologous chromosomes, indicating potential misassemblies or chimeric joins. We averaged the H-trans error across all homologous chromosome pairs to obtain a genome-wide estimate.

Identification of telomeric and centromeric sequences

To identify telomeric and centromeric sequences in the POJ2878 genome, we used quarTeT⁵⁹ (v.1.2.2), using its TeloExplorer and CentroMiner modules for targeted analysis.

Telomeric sequence identification. Using the TeloExplorer module, we first identified the most enriched telomeric-like repeat sequences at the ends of each chromosome. The distribution of these repeats was then analysed to determine the precise location and copy number of telomeric regions on each chromosome.

Centromeric sequence identification. For centromeres, we applied the CentroMiner module to detect tandem repeat monomers. Candidate centromeric sequences were selected on the basis of their repeat period and copy number. To reduce redundancy, the candidate monomers were clustered, and representative monomers were aligned back to the chromosomes. Regions with continuous matches were designated as candidate centromeric regions. These regions were further refined by considering match length and the abundance of retrotransposons, which are characteristic of centromeric regions. Finally, centromere types were classified based on the ratio of the distance between the centromere and the chromosome end (D) to the total chromosome length (L). Chromosomes were categorized as: (1) telocentric if $D/L < 0.001$; (2) acrocentric if $0.001 \leq D/L < 0.1$; (3) submetacentric if $0.1 \leq D/L < 0.4$; and (4) metacentric if $0.4 \leq D/L \leq 0.5$.

Analysis of genome architecture

Subgenome identification. To dissect the distribution of ancestral genetic components in the genome, this study selected ten individuals from each representative ancestral lineage to generate high-quality resequencing data (with sequencing depth of over 10×). Using the POJ2878 genome as a reference, we used the commercial GTX (v.2.1.0; <http://www.gtxlab.com/en/product/cat>) software package for SNP detection and implemented stringent quality control measures. Only SNPs meeting the criteria of alignment quality > 30, base quality > 30 and sequencing depth > 5 were retained, while complex variant types were excluded, leaving only biallelic sites. This high-confidence dataset was used for subsequent analyses.

We identified lineage-specific SNPs by defining *S. officinarum*-specific sites as those where >70% of *S. officinarum* and *S. spontaneum* samples were homozygous for the reference (0/0) and alternate (1/1) alleles, respectively. The reciprocal pattern was used to define *S. spontaneum*-specific SNPs.

We performed a sliding-window analysis with a 50 kb window size and 10 kb step size to examine the distribution of these ancestral components across the genome. We counted the number of *S. officinarum*-specific and *S. spontaneum*-specific sites in each window. A window was classified as belonging to the *S. officinarum* lineage if it contained at least 1.5× more *S. officinarum*-specific sites than *S. spontaneum*-specific sites, and vice versa for the *S. spontaneum* lineage. Regions that did not meet these criteria were labelled as undetermined (UN).

For regions marked as UN, we applied two methods for further classification. The first method used an optimization strategy based on neighbouring information, reflecting the tendency of chromosomal recombination to occur in large, linked segments. If an UN region was flanked by at least three consecutive windows of the same lineage, it was reclassified according to its neighbours. The second method inferred the dominant ancestral affiliation of the entire chromosome. For each chromosome, windows that were not labelled UN and contained at least five *S. officinarum*- or *S. spontaneum*-specific SNPs were considered to be valid. If more than 50% of these valid windows belonged to the *S. officinarum* lineage, the chromosome's dominant ancestry was classified as *S. officinarum*, and any UN regions on that chromosome were

redefined accordingly; the same process applied for the *S. spontaneum* lineage.

Identification of recombined chromosomes. The identification of recombined chromosomes was also based on a sliding window approach with the window size set up to one-tenth of the chromosome length and a step size of 1 Mb. If a chromosome contains two or more windows where regions with lower ancestry proportions exceed 60%, it is classified as a recombined chromosome between two subgenomes.

Identification of HEs. The homoeologous exchanges (HEs) were identified using classified genomic windows (that is, *S. officinarum*- and *S. spontaneum*-assigned windows) as described above. HEs involving a switch from the *S. spontaneum* subgenome to the *S. officinarum* subgenome were detected when two consecutive windows, located within an *S. officinarum*-derived non-recombined chromosome, showed more than 70% of resequenced *S. spontaneum* samples supporting their origin from the *S. spontaneum* species. Similarly, HEs involving a switch from the *S. officinarum* subgenome to the *S. officinarum* subgenome were identified following the same criteria in reverse.

Design of oligo-FISH probes and chromosome painting. Haplotype-specific oligonucleotides were designed against the POJ2878 reference genome using Chorus2 software⁶⁰ with parameters ‘-l 40 -homology 93 -d 5 -step 1’. The initial oligonucleotides were subsequently filtered against whole-genome sequencing data to eliminate sequences with high coverage depth, which may represent potential repetitive sequences. The resulting oligonucleotides were designated as POJ2878 haplotype-specific probes. Chromosome painting by oligo-FISH was performed as described previously⁶¹. Probes were labelled by PCR amplification with Cy3-dUTP or FITC-dUTP (Roche Diagnostics). The hybridization cocktail was composed of 10 µl deionized formamide, 2 µl 20× SSC, 4 µl 50% dextran sulfate and approximately 100 ng of oligonucleotide probes. After denaturation at 95 °C for 10 min, the mixture was applied to denatured chromosome slides and incubated at 37 °C for 16 h, followed by standard washing procedures⁶¹. Chromosomes were counterstained with DAPI (Vector Laboratories) and imaged using an Olympus BX53 fluorescence microscope with a cooled CCD camera, then processed through cellSens Dimension 1.9.

Allele-defined gene annotation and expression

Genome annotation. An integrative approach was used for gene prediction, combining evidence from three sources: ab initio prediction, homologous protein alignments and RNA-seq read mapping. This process was implemented using the GETA program, using clean RNA-seq datasets, homologous proteins and the POJ2878 genome as references. The GETA pipeline used sequencing-depth-based coverage thresholds to filter low-quality transcripts and refine intron predictions. These refined structures were then used to guide the iterative training of gene models in AUGUSTUS⁶² (v.3.3.3), a process that continued until the highest predictive scores were achieved. Homologous proteins from six related plant species (*S. spontaneum*, *Oryza sativa*, *Sorghum bicolor*, *Eriandrus rufipilus*, *Zea mays* and *A. thaliana*)—retrieved from Phytozome and GWH database—were further refined using GENewise⁶³ (v.2.4.1) for protein identification. The final gene models were selected based on sequence conservation, expression levels and AUGUSTUS support scores.

To ensure the reliability of annotations, a stringent filtering process was applied. Genes were aligned against the NR database and the PFAM conserved domain database, and their expression levels were calculated using RNA-seq data. Genes were retained if they met at least one of the following criteria: (1) presence of a homologous protein in the NR database; (2) presence of a conserved domain in the PFAM database; or (3) an FPKM value greater than 3 with sequence coverage exceeding 50%.

Annotation and analysis of TEs. The transposable elements (TEs) of each genome were identified using the EDTA⁶⁴ package using the default parameters. This method provides a whole-genome de novo TE annotation by integrating structural and homology-based approaches, filtering out false positives from raw TE candidates and generating a high-quality, non-redundant TE library for comprehensive genome-wide TE annotations. RepeatMasker (<http://www.repeat-masker.org>) and RepeatModeler2 (ref. 65) were used as the TE libraries for homology-based annotation. Insertion time for full-length LTR-RTs were determined based on the divergence between flanking LTR sequences. Based on the Kimura two-parameter model⁶⁶, we applied the rate⁶⁷ of 6.5×10^{-9} to the formula $T = K/2r$ to convert genetic distance into absolute time.

Identification of allelic and homoeologous genes. To identify genes with defined alleles, paralogous genes and haplotype-specific genes in the complex polyploid sugarcane genome, we developed a three-step method, as illustrated in Extended Data Fig. 3:

Identification of paralogues within each haplotype genome. We constructed the haplotype genome by selecting one chromosome from each homologous chromosome group in the order of haplotype numbering (that is, from g1 to g13). Paralogues within each haplotype genome were identified using hierarchical clustering combined with tree-cutting methods, based on sequence identity and coverage. To ensure accuracy, we dynamically determined the tree-cutting height to achieve paralogue identities exceeding 90% (category 6). Specifically, sequence alignment was performed using DIAMOND⁶⁸ (--query-cover 60 --subject-cover 60 --id 80), generating a gene–gene matrix where each element represented the product of coverage and identity between gene pairs. Hierarchical clustering was then applied to this matrix using Ward’s method with Euclidean distance as the dissimilarity metric⁶⁹. This approach minimized within-cluster variance, ensuring that genes within the same cluster exhibited high similarity while maintaining clear distinctions between clusters. A dendrogram was constructed to represent the relationships among genes, with branch heights reflecting dissimilarity between clusters. Starting from a tree-cutting height of 2, the threshold was gradually reduced by 0.1 until paralogous gene pairs with over 90% identity were identified. This step was performed within each subgenome to minimize interference in subsequent allele identification.

Identification of syntenic and non-syntenic allelic/homoeologous genes. Alleles were identified using JCVI⁷⁰ to compare synteny blocks across haplotype genomes. Each haplotype genome was used as a reference to ensure comprehensive gene representation. Syntenic blocks identified from different haplotype reference genomes were merged, and low-quality syntenic gene pairs were filtered using an identity threshold of 80%. High-quality syntenic gene pairs were classified as syntenic allelic or homoeologous genes (categories 1, 2 and 3). For non-syntenic homologous genes, if the gene exhibited over 90% sequence identity with any classified allelic/homoeologous gene, it was considered as non-syntenic allelic or homoeologous genes (category 4).

Clustering of unclassified genes and identification of single genes. Unclassified genes were further analysed using Broccoli⁷¹, a tool for identifying orthologous protein groups and pairs through phylogenetic and network-based methods. Broccoli performs rapid phylogenetic analysis, builds orthologous relationship networks and applies machine learning to classify groups. Using Broccoli, those unclassified genes were categorized as either non-syntenic allelic/homoeologous gene pairs (category 5) or haplotype-specific genes (category 7).

This three-step approach enabled the classification of genes into seven categories: categories 1–5 represent genes with defined alleles, category 6 includes paralogous genes and category 7 identifies haplotype-specific genes (Supplementary Table 9).

Allele-specific expression analysis using MAS-ISO-seq. To investigate allele-specific expression patterns in sugarcane, we developed a robust pipeline integrating MAS-ISO-seq data using the program Allele-Express (Supplementary Note 3). This approach enabled precise quantification of gene expression from individual alleles within the complex sugarcane genome.

Initially, raw MAS-ISO-seq data were processed to generate HiFi reads using the ccs (<https://github.com/PacificBiosciences/ccs>) tool. Stringent quality-control measures were applied, filtering reads based on minimum pass number, length, signal-to-noise ratio and quality score. These high-quality HiFi reads were then segmented using skera (<https://github.com/PacificBiosciences/skera>) split, guided by a custom primer design file, to separate reads originating from different targeted regions. Primer sequences, poly(A) tails and chimeric sequences were subsequently removed using lima (<https://github.com/PacificBiosciences/barcoding>) and isoseq refine (<https://github.com/PacificBiosciences/IsoSeq>), yielding full-length non-chimeric reads for downstream analysis.

To create a non-redundant reference set, similar transcripts were clustered using mmseqs⁷², retaining only one copy of identical sequences. The FLNC reads were then aligned to these clustered reference transcripts using minimap2⁷³, optimized for splice junctions. The sequencing depth at each site was calculated using bedtools⁷⁴, and saturation curves were generated to confirm sufficient sequencing depth.

A key component of our pipeline, Allele-Express, systematically identified and filtered candidate read alignments for each homologous gene group. Reads sharing the same identifier were grouped and a multi-layered filtering strategy was implemented. Initially, a priority alignment was selected on the basis of mapping quality, edit distance and alignment score. For ambiguous alignments, where a read could potentially map to multiple alleles, we developed two complementary approaches:

- (1) Read depth stable point (RDSP) method: this approach identifies the point within a transcript at where the read depth stabilizes, discarding reads aligning beyond this point as ambiguous.
- (2) Read clustering method: this method performs a comprehensive sequence comparison of all full-length MAS-ISO-seq reads using a greedy incremental clustering algorithm. Ambiguous reads are assigned to a specific transcript if other reads within the same cluster align uniquely to that transcript.

Finally, Allele-Express quantified read counts for each transcript and aggregated them to the gene level using transcript-to-gene mappings. Read counts were normalized using TPM to enable comparisons across samples. This integrated pipeline, combining MAS-ISO-seq and Allele-Express, provides a comprehensive analysis of allele-specific expression in sugarcane, offering valuable insights into the regulatory mechanisms governing this complex polyploid genome.

Haplotype counts were determined using allele-aware annotations derived from the haplotype-resolved genome assembly. Although multiple copies of a gene may be annotated, only haplotypes containing at least one mutation, compared with alternative forms, were included in the analysis. For example, while a total of 12 copies of the *MNS2* gene were annotated, only eight haplotypes were retained for further analysis: three allelic genes from *S. spontaneum* and five from *S. officinarum*.

ASHEGs were identified using the following criteria: within each allele group, the allele with the highest TPM must exhibit an expression level at least 2.5× greater than the second-highest allele. A Grubbs test was conducted to identify outliers, with a $P < 0.05$ indicating statistical significance. The allele with the highest TPM was considered a candidate if it qualified as an outlier. The second-highest allele in the group must not be classified as an outlier.

Population genetics

Whole-genome resequencing and variant detection. Sequencing data were aligned to the reference genome POJ2878 using GTX (v2.1.0), ensuring unique alignments by filtering out 'SA' and 'XA' tags in BAM files and retaining reads with alignment scores $AS:i > XS:i$ and mapping quality $mapq > 20$. Variants were called using HaplotypeCaller tool, and SNP filtering was performed with bcftools⁷⁵ (v.1.10.2) using stringent criteria: $QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 || SOR > 3.0$. Additional filtering retained only biallelic SNPs with minor allele frequency (MAF) ≥ 0.05 and a missing rate $\leq 50\%$. A total of 11,427,420 high-quality SNPs was identified for downstream analysis. SNP annotation was performed using SnpEff⁷⁶ (v.5.2c). Among these SNPs, 656,089 (5.74%) were shared between *S. spontaneum* and *S. officinarum* in the POJ2878 genome (defined as sites with call rate $> 50\%$ in both populations). To assess whether shared loci correspond to conserved homologous sequence contexts across progenitor genomes, 500 loci were randomly sampled from the 656,089 shared set and evaluated for syntenic context using the T2T monoploid assemblies of *S. spontaneum* 82-114 and *S. officinarum* XZ. A locus was deemed to be conserved if its SNP coordinate and ± 200 bp flanking sequence were identifiable in POJ2878, *S. spontaneum* 82-114 and *S. officinarum* XZ within syntenic regions. Based on this criterion, 57.8% (289 out of 500) of loci were conserved across all three genomes, while 27.2% (56 in *S. officinarum* and 80 in *S. spontaneum*) were detected in at least one progenitor genome. The remaining unmatched loci probably reflect local assembly gaps, structural or presence-absence variation among genomes, and/or stringent matching thresholds, rather than mapping artefacts.

Phylogenetic analysis and population structure. Phylogenetic relationships were inferred using VCF2Dis (v.1.52; <https://github.com/BGI-shenzhen/VCF2Dis>) to construct a neighbour-joining tree. Population structure was analysed with Admixture⁷⁷ (v.1.3.0). Linkage disequilibrium (LD) decay was calculated and visualized using PopLDdecay⁷⁸ (v.3.42), and PCA was conducted using VCF2PCAcluster⁷⁹ (v.1.40).

Identification of shared IBD regions and favourable haplotypes. To investigate shared IBD regions across sugarcane accessions, we used IBDseq⁸⁰ (r1206) with its default settings. We quantified the contribution of each IBD region by introducing a metric, termed IBD density, which measures the proportion of varieties sharing IBD segments with POJ2878. The formula for calculating IBD density is:

$$\text{IBD density} = \frac{\text{Number of varieties sharing IBD segments with POJ2878}}{\text{Total number of varieties examined in the population}}$$

The genome was analysed in 50-kb non-overlapping windows, and regions within the top 5% of IBD density were designated as hotspots of POJ2878 contributions, representing favourable haplotypes. These hotspots mark genomic regions with substantial genetic contributions from POJ2878, and probably harbour key alleles associated with important agronomic traits in sugarcane.

Selective sweep analysis. Selective sweeps within populations were identified using three complementary methods: SweeD⁸¹ (v.4.0.0), RAiSD⁸² (v.2.9) and nucleotide diversity (π). Genomic regions that ranked in the top 5% by any of these methods were considered to be candidate selective sweep regions. Regions identified by at least two methods were classified as high-confidence selective sweeps, providing robust evidence of natural selection. To detect interpopulation selective signals, F_{ST} values and XP-CLR scores were calculated using VCFtools⁸³ (v.0.1.16) and XP-CLR (<https://github.com/hardingnj/xpclr>), respectively. Domestication-related selective sweeps in *S. officinarum*

were identified by comparing it to 43 resequencing individuals of *S. robustum*, downloaded from a previous publication³. Moreover, hybrid sugarcanes were compared against both *S. spontaneum* and *S. officinarum* to detect selective signals specific to hybrid populations. The thresholds for selective sweeps were set at the top 5% for hybrids and the top 5% for *S. officinarum*. All analyses used a sliding-window approach with 50-kb window size and 10-kb step size. Regions supported by both F_{ST} and XP-CLR were designated as high-confidence selective sweeps, highlighting key genomic signatures linked to domestication and hybridization.

Application of KMERIA in Sugarcane GWAS

Traditional SNP-based GWAS face great challenges when applied to polyploid genomes. These include genotyping errors due to ambiguous read alignments, reliance on biallelic variants despite the prevalence of multiallelic variants in polyploid genomes and the time-intensive nature of the process. To address these challenges, we applied a k -mer based GWAS approach (KMERIA), which has been submitted elsewhere as a companion paper¹⁵. This method leverages short, fixed-length sequence subsequences (k -mers) directly from resequencing data, eliminating the need for reference genome. The algorithm package has key steps. k -mers are counted from whole-genome sequencing reads of all samples using KMERIA with a recommended k -mer size of 31 bp for optimal efficiency. To improve accuracy, k -mers appearing once in a single individual are filtered out. A population k -mer matrix is built from individual k -mer counts using the KMERIA kctm function. Subsequently, k -mers at both very low and high frequencies (representing sequencing errors or repetitive regions) are filtered based on the genome ploidy and k -mer frequency distribution for each individual. k -mers exceeding a predefined individual missing rate are removed to ensure valid population-level associations. Individual sequencing depths are incorporated to account for their influence on k -mer copy number statistics.

To calculate the allele dosage, k -mer abundance values are transformed to better suit the GWAS model and enhance allele dosage estimation. The k -mer abundance below the 5th percentile is set to 0, and values exceeding the 95th percentile are set to 2. The remaining values are linearly transformed between 0 and 2 based on the original k -mer abundance value.

To perform GWAS using the transformed k -mers, we estimate the population structure and kinship relatedness through randomly selecting a small proportion (approximately 0.1%) of recoded k -mer genotypes. The association test further uses a mixed linear model incorporating a series of factors, including fixed effect, genotyped k -mers principal component vector, polygenic random effect vector and residual error.

To achieve precise genomic localization, KMERIA uses a read-based refinement strategy. Instead of mapping k -mers directly, we first retrieved sequencing reads that contain the significant k -mers. These longer, context-rich reads are then searched against the monoploid reference genome by BWA (default parameters) or minimap2 (option -x sr). This approach substantially improves alignment accuracy and enables unique localization of the associated genomic regions with high confidence. Once significant k -mers are anchored through this read-based mapping, KMERIA defines a candidate genomic interval for gene discovery. By default, a 50 kb window is surveyed on both sides of the mapped site, although this can be adjusted according to the LD decay of the studied population. All annotated genes within this interval are considered putative candidates underlying the associated trait.

To distinguish the most likely causal genes from the initial candidates, we integrated statistical causal inference with functional evidence. Fine-mapping was performed within the ± 50 kb genomic windows centred on the significantly associated k -mers. We used the CAVIAR⁸⁴ to quantify the causal posterior probability (CPP) for each

associated k -mer within the interval by integrating GWAS z -scores with a local LD correlation matrix calculated using PLINK2 using the --r-unphased square option⁸⁵. CAVIAR also identifies a minimal causal set, representing the smallest group of k -mers with a collective confidence level of $\geq 95\%$ in capturing all true causal variants. Within this set, the top five k -mers with the highest CPP were designated as the primary statistical anchors for gene prioritization. To ensure the biological relevance of the statistically identified candidates, genes within the fine-mapped intervals were prioritized based on the following criteria: (1) priority was given to genes that either overlapped with the k -mer exhibiting the highest CPP or were located within its immediate flanking regions. (2) Candidates were referenced for functional orthologues with genes previously reported to be involved in biosynthesis or regulatory pathways in related species (such as *S. bicolor*, *O. sativa* and *Z. mays*). This integrated approach facilitated the transition from broad association intervals to high-confidence functional candidates.

Association analysis with parenchyma cell traits and Brix

Tissue sections of sugarcane stems. The experiment was conducted using different sugarcane cultivars grown in Lingao County, Hainan Province, China. The sampling was performed in September 2024, during the pre-mature growth stage, approximately 210 days after planting. The samples were collected from the sixth internode above the ground of the sugarcane stem.

Fresh sugarcane stem segments (5 mm $\times$ 3 cm) were fixed in 70% FAA (>48 h) and dehydrated through a graded ethanol series (60–95%; 24–72 h per step), followed by two rounds of 100% ethanol (12–24 h each), before resin embedding. Subsequently, dehydrated samples were infiltrated with T7200 photosensitive resin (KULZER, Technovit7200VLC) for 15–25 days, depending on the sample size. To prevent premature resin polymerization, the infiltration process was conducted in the dark. The infiltrated samples were placed in embedding moulds, and T7200 resin was added to cover the samples. Polymerization was carried out in a photocuring embedding machine (EXAKT, E520) for 12 h. The resin-embedded samples were glued to glass slides (Servicebio) using T4000 adhesive (KULZER, Technovit4000). The tissue surface was ground to a smooth and uniform plane using a hard tissue grinding machine (EXAKT, E400CS) with 1200-grit sandpaper. The samples were sectioned into 200 μ m slices (EXAKT, E300CP) and polished to a final thickness of 20–30 μ m. The sugarcane stem tissue sections were stained using the toluidine blue staining protocol as follows. The sections were immersed in toluidine blue staining solution (Servicebio, G1032) for 5 min. The degree of staining was monitored under a Nikon Eclipse E100 upright optical microscope. Overstained sections were differentiated with 95% ethanol. After staining, the sections were rinsed with tap water for 10 s and dried in an oven at 60 °C. The stained sections were cleared in xylene (SCRC, 10023418) for 30 s and mounted with neutral gum (SCRC, 10004160). The stained sections were observed under a Nikon DS-U3 imaging system. A blue-green hue characterized the lignified cell walls, distinct from the purple-blue of the cellulose walls.

Automated parenchyma cell analysis in sugarcane stems. To quantify parenchyma cell density and size in sugarcane stems, we developed an automated image analysis pipeline for high-resolution scanned virtual slides (SVS). This workflow comprises three key steps.

Image tiling and conversion: the high-resolution SVS files were systematically divided into non-overlapping tiles (2,560 $\times$ 2,560 pixels) and converted to a standard image format for efficient processing.

Background filtering and vascular bundle identification: to isolate parenchyma tissue, we first removed background areas. Images were divided into blocks (320 $\times$ 320 pixels), and large white blocks were identified as background and excluded. The remaining blocks were analysed to

identify vascular bundles based on their distinct staining characteristics after Toluidine Blue treatment and their morphology.

This process involved further dividing the non-background image into smaller blocks (160 × 160 pixels) and calculating the average RGB colour values for each block, creating a colour feature matrix. Toluidine Blue staining creates a strong colour contrast between parenchyma cells and the sieve tube portion of vascular bundles. However, the xylem tissue within the bundles shares similar colour and morphological features with parenchyma cells, making differentiation challenging. To address this, a Gaussian smoothing filter was applied to the colour feature matrix, reducing noise and enhancing colour transitions. This smoothing process effectively corrected the xylem tissue colour to match the rest of the vascular bundle and also reduced dye-induced noise in the parenchyma regions. Finally, K-means clustering was applied to the smoothed colour feature matrix to separate vascular bundles and parenchyma cell regions into two distinct groups.

For cell segmentation and counting, parenchyma cells within the identified regions were segmented using Cellpose⁸⁶, using the cyto2 model with optimized parameters (minimum cell area, 6,400 pixels; cell probability threshold, 2) to ensure accurate detection. The number of parenchyma cells was then quantified based on these segmented regions.

This automated method enables precise identification and quantification of parenchyma cells in sugarcane stem cross-sections. To validate its accuracy, we compared automated counts and size measurements with manual counts from ten randomly selected stem cross-sections. A strong positive correlation was observed for both cell density (Pearson's $r = 0.9828$, $P < 0.0001$) and cell size (Pearson's $r = 0.9155$, $P = 0.0002$), demonstrating the reliability of our automated approach (Extended Data Fig. 7).

Assessment of Brix. For each sugarcane variety, juice was extracted from at least six mature stalks. The soluble solids content was measured using a digital refractometer (PAL-1, ATAGO). The refractometer was calibrated with distilled water and adjusted to 28 °C. The SSC values were expressed in degrees Brix.

Genetic transformation in the hybrid sugarcane cultivar

Vector construction and activation of *Agrobacterium* strains. The open reading frame sequence of ShSUT2, ShTIP1, ShCBL1 and ShATG18a were amplified by PCR and cloned into the downstream region of the maize *Ubi* promoter in the pCambia3300 vector through homologous recombination. The construct was transferred into *Agrobacterium tumefaciens* EHA105 through freeze–thaw transformation. Before transformation, the positive clones were activated twice on YEP agar plates. The bacterial cells were then collected into the infection medium, incubated at 28 °C with agitation (90 rpm) for 60–90 min to achieve an optical density at 600 nm of 0.3–0.6, and subsequently used for callus infection and genetic transformation.

Embryogenic callus induction and transformation. The young leave rolls from the stem apex of sugarcane cultivar XTT22 were sterilized using ethanol and mercury chloride (HgCl₂), and then cut into 1–2-mm-thick sections. These sections were cultured on an induction medium (MS + 30 g l⁻¹ sucrose + 2 mg l⁻¹ 2,4-D) to induce embryogenic callus formation at 28 °C in darkness, changing the medium twice during the induction period until embryogenic calli emerged. Approximately 5 g of embryogenic calli was collected and incubated in *Agrobacterium* infection solution under agitation (90 rpm) at 28 °C for 30 min. Inoculated calli were dried and incubated at 22 °C in darkness for 3 days. Subsequently, they were transferred to recovery medium (MS + 30 g l⁻¹ sucrose + 2 mg l⁻¹ 2,4-D) for 1 week, followed by selection on screening medium (MS + 30 g l⁻¹ sucrose + 2 mg l⁻¹ 2,4-D + 2 mg l⁻¹ Basta + 300 mg l⁻¹ Timentin) for 25–30 days. Surviving healthy embryogenic calli were selected to culture on differentiation

medium (MS + 30 g l⁻¹ sucrose + 1.5 mg l⁻¹ 6-BA + 2 mg l⁻¹ Basta + 300 mg l⁻¹ Timentin) to regenerate shoots, which were further rooted on rooting medium (MS + 30 g l⁻¹ sucrose + 1.5 mg l⁻¹ NAA + 2 mg l⁻¹ Basta + 300 mg l⁻¹ Timentin). Finally, the rooted seedlings were transplanted into plug trays for growth. Transgenic lines were verified using a PAT/bar lateral flow assay test kit (Genesino) using aqueous leaf homogenates (1:20, w/v).

Statistics and reproducibility. All experiments were repeated independently at least three times with similar results. This includes all assays from which representative images are shown, such as micrographs of plant tissue sections, plant phenotypes and gels.

Reporting summary

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

Data availability

The raw sequencing data have been deposited in NCBI database under accession number of PRJNA1303125. The genome assembly and annotation are available at Figshare⁸⁸ (<https://figshare.com/s/e79249d6ce11539fa105>) and NCBI database under accession number PRJNA1312687. Source data are provided with this paper.

Code availability

The codes of C-Phasing, Allele-Express and KMERIA have been released in GitHub and Zenodo⁸⁷ (<https://doi.org/10.5281/zenodo.18544719>); CPhasing (<https://github.com/wangyibin/CPhasing>); Allele-Express (<https://github.com/xiaochen-eden/Allele-Express>); KMERIA (<https://github.com/ShIne111/KMERIA>). The codes for automated analysis of parenchyma cell have been archived at GitHub (<https://github.com/Chaiguanggun/Automated-Parenchyma-Cell-Analysis-in-Sugarcane-Stems>). The codes for the identification of allelic or homoeologous genes are available at GitHub (<https://github.com/Chaiguanggun/PolyHomology>).

46. Cheng, H., Asri, M., Lucas, J., Koren, S. & Li, H. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. *Nat. Methods* **21**, 967–970 (2024).
47. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).
48. Zeng, X. et al. Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nat. Plants* **10**, 1184–1200 (2024).
49. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).
50. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
51. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
52. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
53. Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mobile DNA* **10**, 48 (2019).
54. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
55. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR assembly index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
56. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421 (2009).
57. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
58. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <https://arxiv.org/abs/1303.3997v2> (2013).
59. Lin, Y. et al. quarteT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic. Res.* **10**, uhad127 (2023).
60. Zhang, T., Liu, G., Zhao, H., Braz, G. T. & Jiang, J. Chorus2: design of genome-scale oligonucleotide-based probes for fluorescence in situ hybridization. *Plant Biotechnol. J.* **19**, 1967–1978 (2021).

61. Yu, F. et al. Chromosome-specific painting unveils chromosomal fusions and distinct allopolyploid species in the *Saccharum* complex. *N. Phytol.* **233**, 1953–1965 (2022).
62. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312 (2004).
63. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
64. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
65. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457 (2020).
66. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120 (1980).
67. Gaut, B. S., Morton, B. R., McCaig, B. C. & Clegg, M. T. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcl*. *Proc. Natl Acad. Sci. USA* **93**, 10274–10279 (1996).
68. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
69. Ward, J. H. Jr. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58**, 236–244 (1963).
70. Tang, H. et al. JCVI: a versatile toolkit for comparative genomics analysis. *iMeta* **3**, e211 (2024).
71. Derelle, R., Philippe, H. & Colbourne, J. K. Broccoli: combining phylogenetic and network analyses for orthology assignment. *Mol. Biol. Evol.* **37**, 3389–3396 (2020).
72. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
73. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
74. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
75. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
76. Cingolani, P. in *Variant Calling: Methods and Protocols* (eds Ng, C. & Pisuoglio, S.) 289–314 (Springer, 2022).
77. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
78. Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M. & Yang, T.-L. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786–1788 (2019).
79. He, W. et al. VCF2PCACluster: a simple, fast and memory-efficient tool for principal component analysis of tens of millions of SNPs. *BMC Bioinform.* **25**, 173 (2024).
80. Browning, B. L. & Browning, S. R. Detecting identity by descent and estimating genotype error rates in sequence data. *Am. J. Hum. Genet.* **93**, 840–851 (2013).
81. Pavlidis, P., Živković, D., Stamatakis, A. & Alachiotis, N. Sweed: likelihood-based detection of selective sweeps in thousands of genomes. *Mol. Biol. Evol.* **30**, 2224–2234 (2013).
82. Alachiotis, N. & Pavlidis, P. RAISD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. *Commun. Biol.* **1**, 79 (2018).
83. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
84. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasiński, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).
85. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
86. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106 (2021).
87. Wang, Y., Chen, S. & Chen, X. Custom codes in the POJ2878 genome paper. *Zenodo* <https://doi.org/10.5281/zenodo.18544719> (2026).
88. Wang, J. et al. Fully haplotype-phased genome assembly of *Saccharum* hybrid POJ2878. *Figshare* <https://figshare.com/s/e79249d6ce11539fa105> (2025).

Acknowledgements We thank S. Huang and H. Wang for discussions on this work. This work was supported by National Natural Science Foundation of China grant (32222019, 32370714 and 32572476), Basic Research Center for Agricultural Frontiers and Interdisciplinary Sciences (BRC-AFIS), Innovation Program of Chinese Academy of Agricultural Sciences (CAAS-BRC-AFIS-2025-02), the Project of State Key Laboratory for Tropical Crop Breeding (NKLTCBCTXD23, NKLTCB-ZX03, NKLTCB202313), the Key R&D Program of Hainan Province (ZDYF2025XDNY104, ZDYF2025GXJS139), the Central Public-interest Scientific Institution Basal Research Fund (1630052023015), Chinese Academy of Tropical Agricultural Sciences for Science and Technology Innovation Team of National Tropical Agricultural Science Center (CATASCTD202307).

Author contributions X. Zhang, R.M., M.Z. and J.W. designed the project and coordinated research activities. Y.W. developed the C-Phasing algorithm. X.C. developed the Allele-Express algorithm. S.C. developed the KMERIA algorithm. Ping Zhao generated the Pore-C data. Y.W., A.D., Y.L., Q.D. and X. Zeng assembled and assessed genome assembly quality. J.W., Y.W., A.D., Y. Zheng, Z. Ye, S.Q., Y.S., Shengcheng Zhang, R.M., M.Z. and X. Zhang analysed genome architecture and subgenome dynamics. X. Li, J.W. and Z.X. performed population genetics analysis. J.W., J.L. and K.C. generated and analysed parenchyma cell and sugar content data. X. Li, J.W., H.L., T.Z., C.F., R.W., Y. Zhang, R.K., B.Z., W.Z., Peifang Zhao, G.Y., H.J., X. Zhao, L.L., B.C., J.M., X. Lu and Y.G. cultivated plant materials and conducted phenotyping. S.H., Y.H., Z. Yu and Z.D. performed Oligo-FISH experiments. Q.Z., Shuzhen Zhang, W.C. and D.H. conducted the construction of overexpression vectors, genetic transformation and phenotyping in both rice and sugarcane. X. Zhang drafted the manuscript and R.M. and M.Z. revised it. All of the authors read and approved the final manuscript.

Competing interests The authors declare no competing interests.

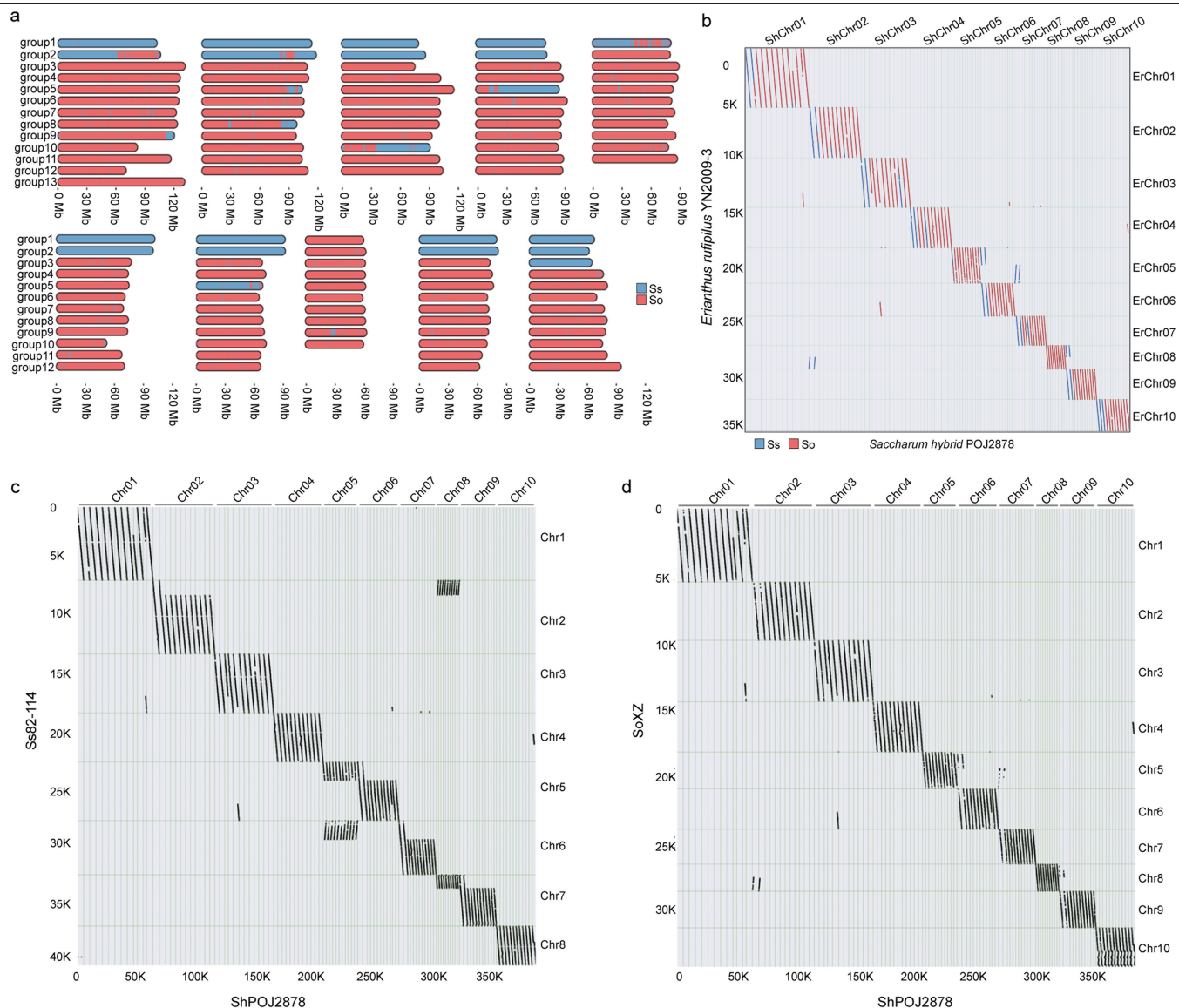
Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-026-10576-7>.

Correspondence and requests for materials should be addressed to Muqing Zhang, Ray Ming or Xingtang Zhang.

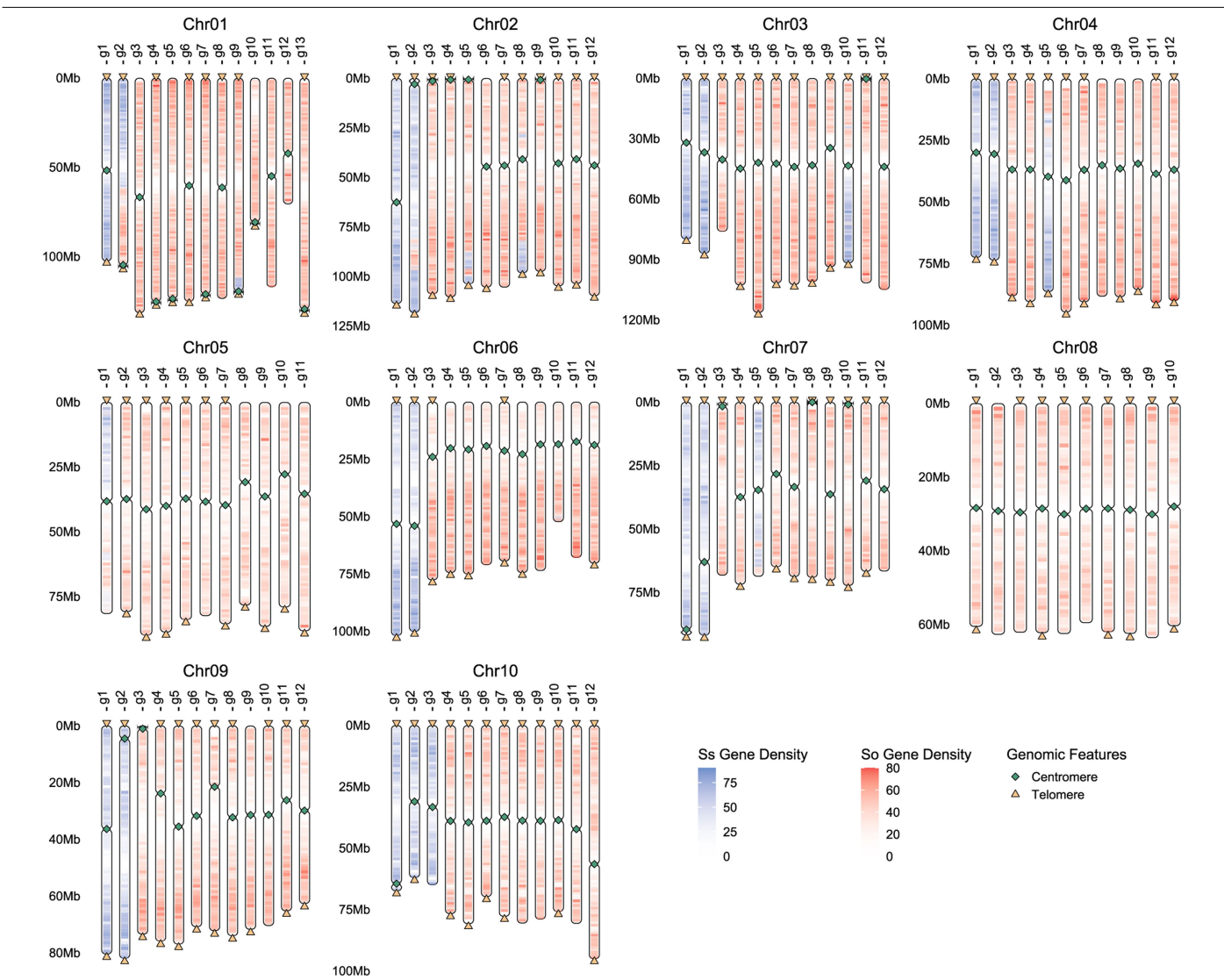
Peer review information *Nature* thanks James Schnable and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



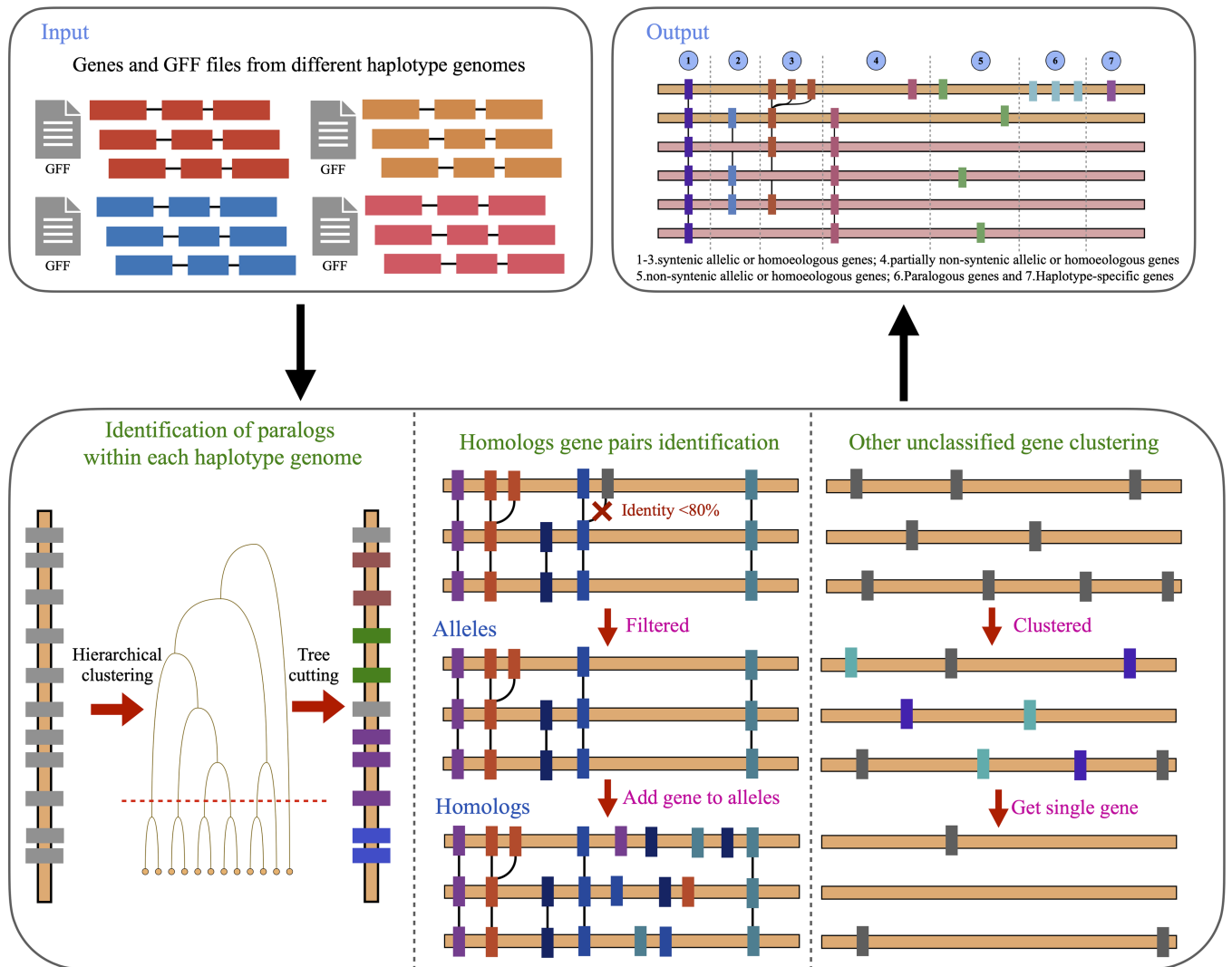
Extended Data Fig. 1 | Subgenome identification and synteny analysis.
(a) Chromosome length and subgenome assignment of the 118 chromosomes in POJ2878, with blue representing Ss-derived sequences and red representing

So-derived sequences. **(b-d)** Dotplot analysis comparing POJ2878 with *Erianthus rufipilus* **(b)**, Ss82-114 **(c)** and SoXZ **(d)**, illustrating chromosomal synteny.



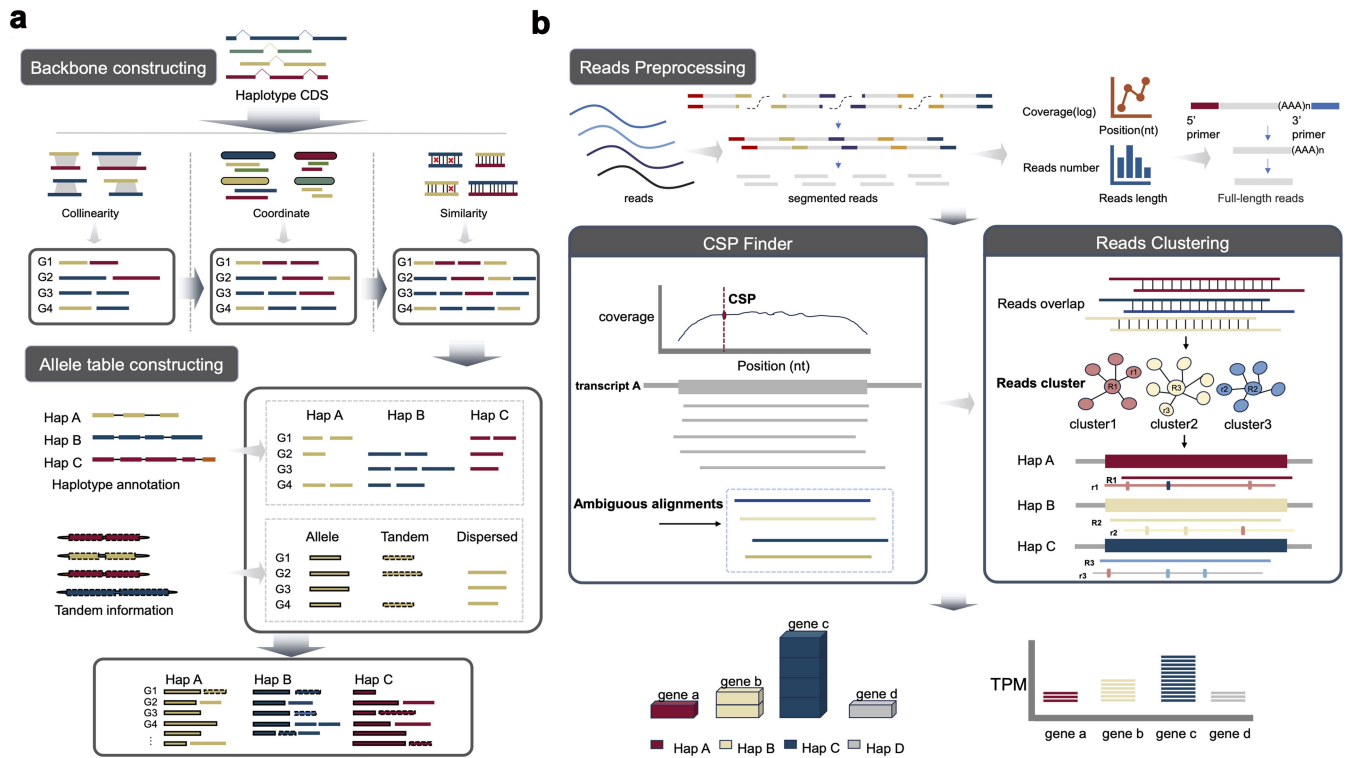
Extended Data Fig. 2 | Genomic features along the 118 chromosomes in POJ2878 genome. Each vertical bar represents an individual chromosome, with its physical length indicated in megabases (Mb) on the y-axis. The density of genes from *S. spontaneum* (Ss) and *S. officinarum* (So) is represented by

heatmaps in blue and red, respectively. The colour intensity corresponds to the gene count within a sliding window. The predicted locations of centromeres and telomeres are marked by green diamonds and yellow triangles, respectively.



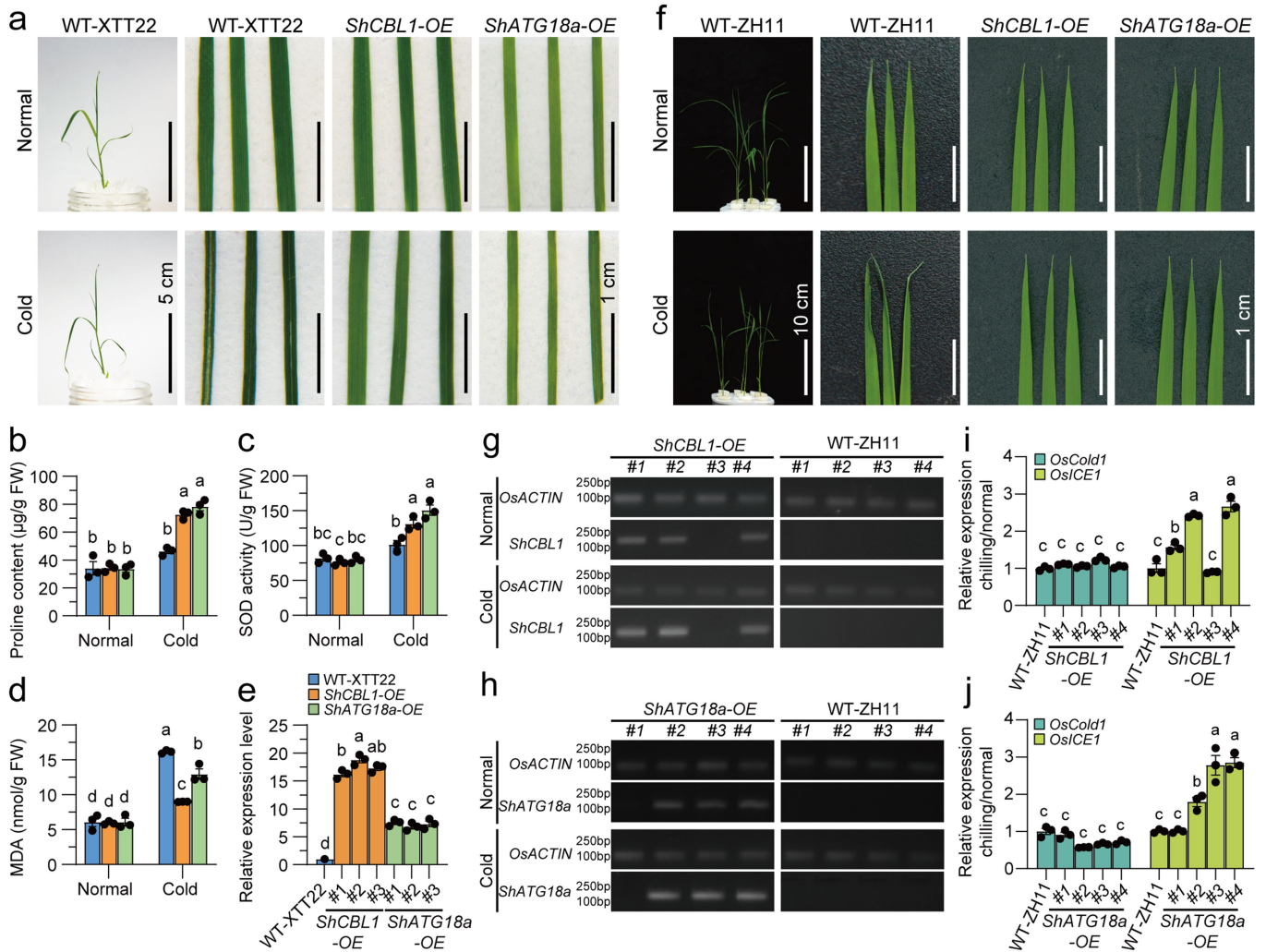
Extended Data Fig. 3 | Schematic of the gene classification pipeline. The pipeline takes gene annotation files (GFF) from different haplotype genomes as input. In the first step, paralogous genes (category 6) are identified within each haplotype using hierarchical clustering and dynamic tree-cutting. The second step identifies allelic/homoeologous genes through cross-haplotype synteny analysis. Syntenic pairs are filtered based on sequence identity (>80%) to define syntenic allelic/homoeologous genes (categories 1-3), while

non-syntenic genes with high similarity to classified genes are assigned as partially non-syntenic allelic/homoeologous genes (category 4). In the third step, remaining unclassified genes are clustered to distinguish non-syntenic allelic/homoeologous genes (category 5) from haplotype-specific genes (category 7). Finally, this workflow comprehensively classifies all genes into seven distinct categories.



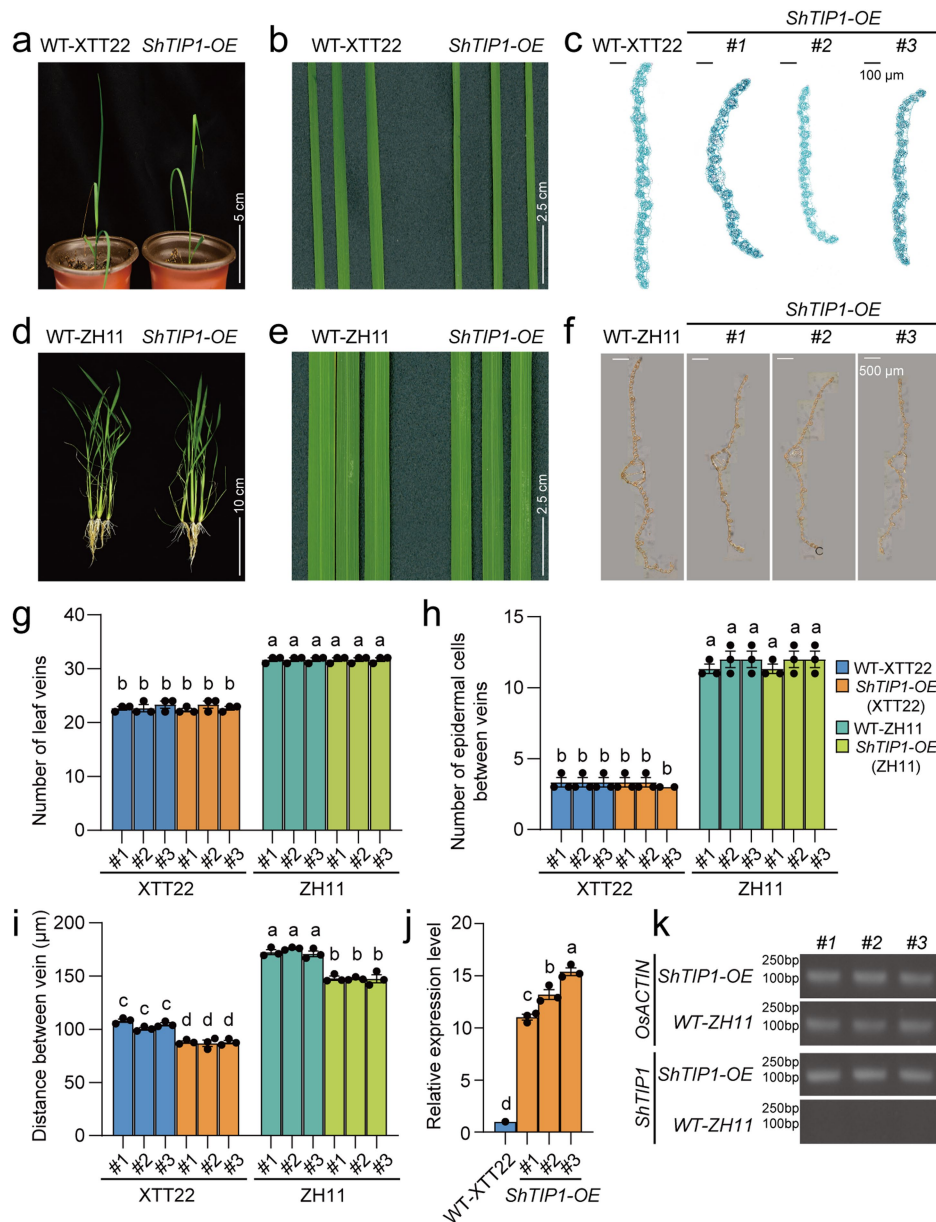
Extended Data Fig. 4 | Overview of the Allele-Express framework. (a) The *Allele* module for identifying allelic genes in haplotype-resolved genomes. The process begins with the construction of a haplotype backbone from Coding Sequences (CDS). An allele table is subsequently constructed by integrating three criteria: collinearity, genomic coordinates, and sequence similarity. This multi-faceted approach classifies genes into allelic, tandem, or dispersed groups, leveraging tandem gene annotations and haplotype information. **(b)** The Expression module for measuring allele-specific gene expression.

This module first preprocesses raw MAS-ISO-seq reads. To resolve ambiguous alignments, two complementary methods are employed: the CSP (coverage stable point) finder, which identifies a stable point in read coverage to trim ambiguous read ends, and a read clustering approach, which groups reads based on sequence similarity to confidently assign ambiguous reads. Finally, the expression of each allele is quantified and normalized to Transcripts Per Kilobase Million (TPM).



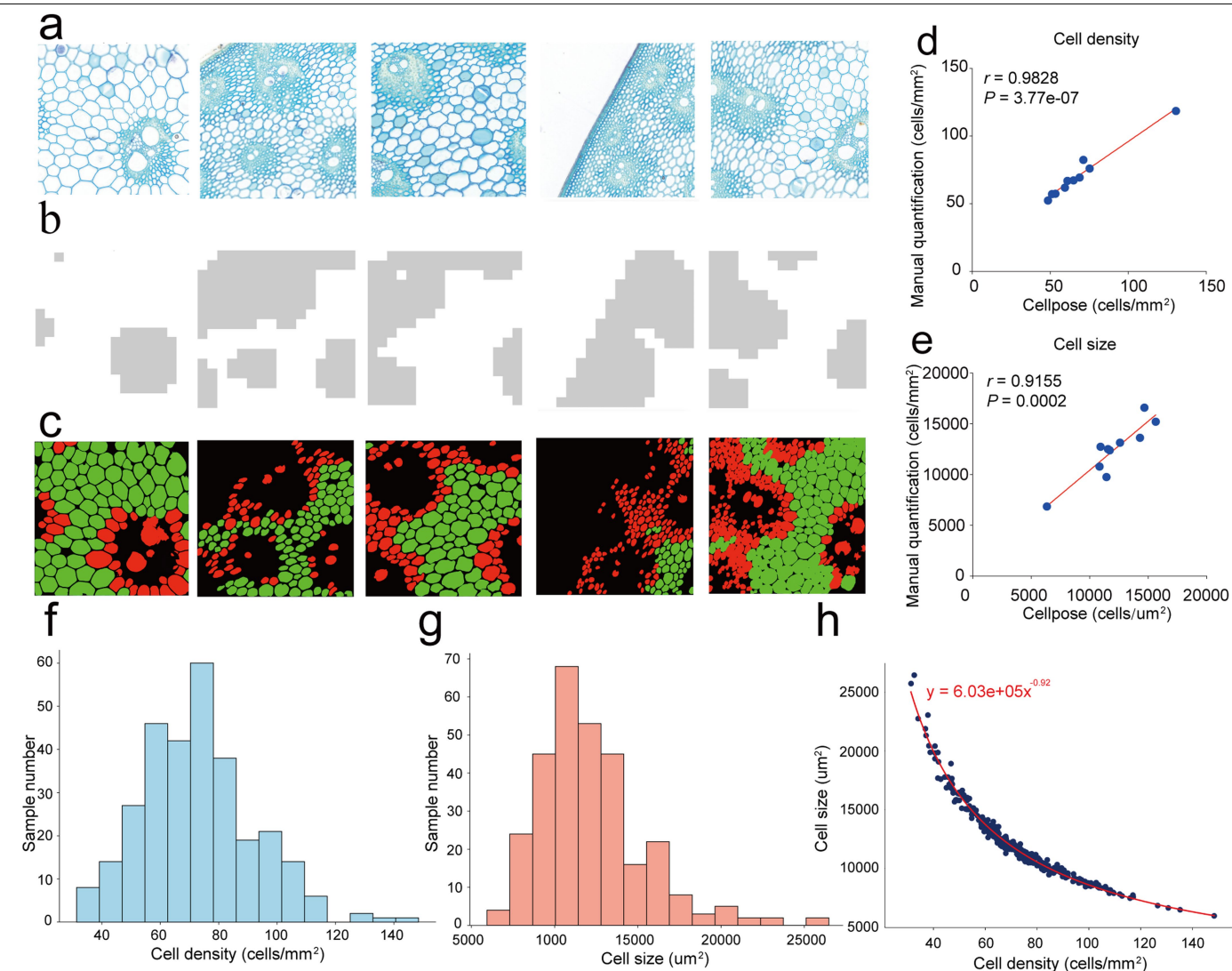
Extended Data Fig. 5 | Overexpression of *ShCBL1* and *ShATG18a* enhances cold tolerance in sugarcane and rice. (a) Phenotypes of wild-type (WT-XTT22) and transgenic sugarcane lines overexpressing *ShCBL1* (*ShCBL1*-OE) or *ShATG18a* (*ShATG18a*-OE) under normal (26 °C) and cold (−5 °C) stress conditions for 7 h. Scale bars = 5 cm (left) and 1 cm (right). (b–d) Measurement of proline content (b), superoxide dismutase (SOD) activity (c), and malondialdehyde (MDA) content (d) in the leaves of WT-XTT22, *ShCBL1*-OE, and *ShATG18a*-OE sugarcane lines under normal and cold conditions. Different letters indicate statistically significant differences ($P < 0.05$, one-way ANOVA). (e) Relative expression levels of *ShCBL1* and *ShATG18a* in WT-XTT22 and transgenic sugarcane lines as determined by qRT-PCR. (f) Phenotypes of wild-type rice (WT-ZH11) and

transgenic rice lines overexpressing *ShCBL1* or *ShATG18a* under normal and cold (4 °C) stress conditions for 5 days. Scale bars = 10 cm (left) and 1 cm (right). (g, h) Semi-quantitative RT-PCR analysis of *ShCBL1* and *ShATG18a* expression in transgenic rice lines under normal and cold conditions (For gel source data, see Supplementary Fig. 1). *OsACTIN* was used as an internal control. (i, j) Relative expression levels of cold-responsive genes *COLD1* and *ICE1* in WT-ZH11, *ShCBL1*-OE (i), and *ShATG18a*-OE (j) rice lines under normal and cold conditions. For b, c, d, e, i, and j, data are presented as mean $\pm$ SD ($n = 3$ independent biological replicates per genotype). Different letters indicate statistically significant differences ($P < 0.05$, one-way ANOVA). Exact P values are provided in the Source data.



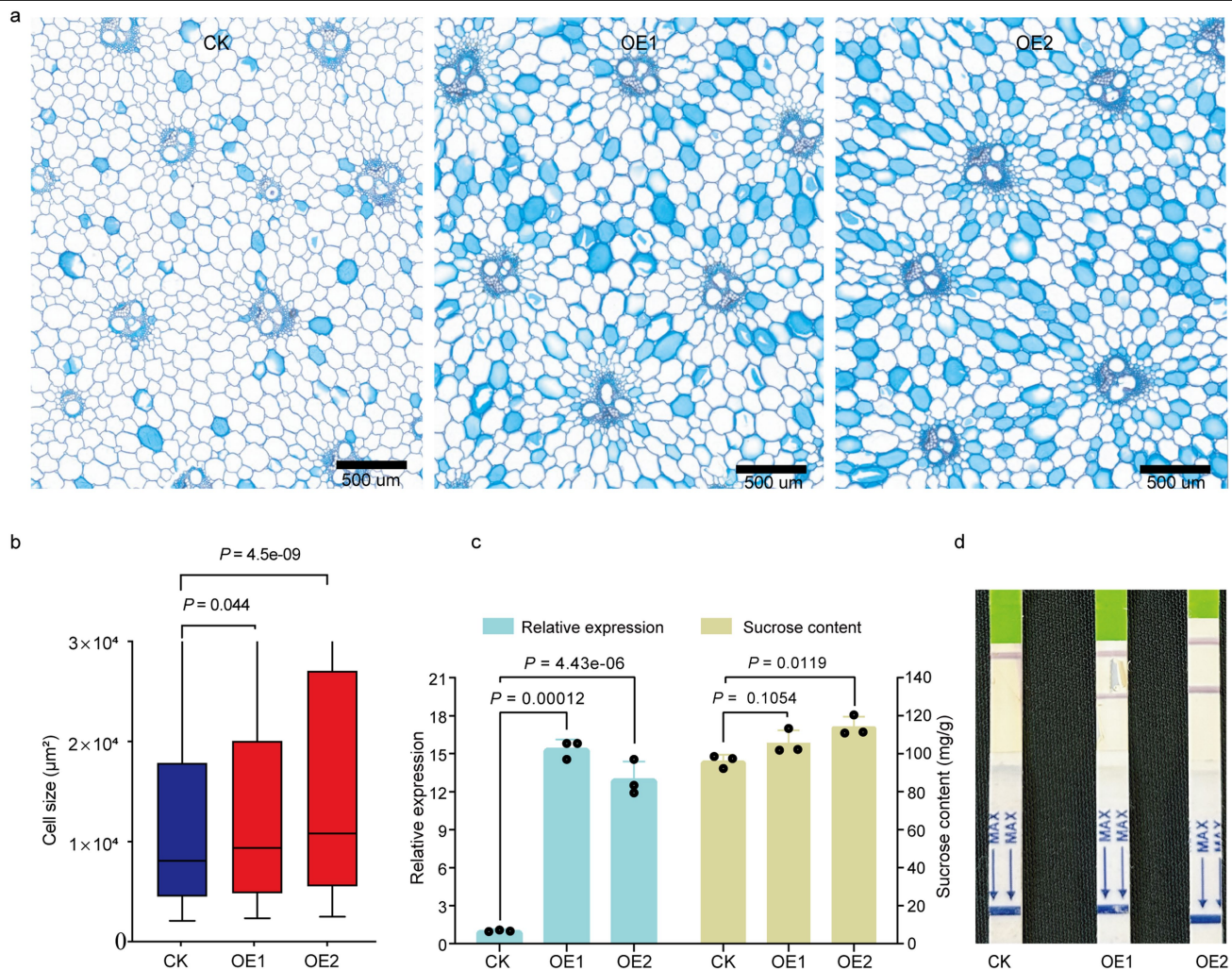
Extended Data Fig. 6 | Overexpression of *ShTIP1* alters leaf morphology in sugarcane and rice. (a) Phenotypes of 30-day-old wild-type (WT-XTT22) and *ShTIP1*-overexpressing (*ShTIP1*-OE) sugarcane plants. Scale bar, 5 cm. (b) Leaf morphology of WT-XTT22 and *ShTIP1*-OE sugarcane. Scale bar, 2.5 cm. (c) Transverse sections from the middle of leaves of WT-XTT22 and three independent *ShTIP1*-OE transgenic sugarcane lines (#1, #2, #3). Scale bar, 100 μm. (d) Phenotypes of 60-day-old WT (WT-ZH11) and *ShTIP1*-OE rice plants. Scale bar, 10 cm. (e) Leaf morphology of WT-ZH11 and *ShTIP1*-OE rice. Scale bar, 5 cm. (f) Transverse sections from the middle of leaves of WT-ZH11 and three independent *ShTIP1*-OE transgenic rice lines (#1, #2, #3). Scale bar, 500 μm. (g-i) Quantification of leaf anatomical traits in WT and *ShTIP1*-OE lines of

sugarcane (XTT22) and rice (ZH11), including (g) number of leaf veins, (h) number of epidermal cells between veins, and (i) distance between veins. (j) Relative expression analysis of *ShTIP1*, in both WT and *ShTIP1*-OE lines of sugarcane and rice. In panels (g-j), data are presented as mean ± SD ($n = 3$ independent biological replicates). Different letters (a, b, c, d) above the bars indicate statistically significant differences between the lines, as determined by a one-way ANOVA test ($P < 0.05$). Exact P values are provided in the Source data. (k) Semi-quantitative RT-PCR analysis of *ShTIP1* expression in transgenic rice lines with *OsACTIN* as an internal control (For gel source data, see Supplementary Fig. 1).



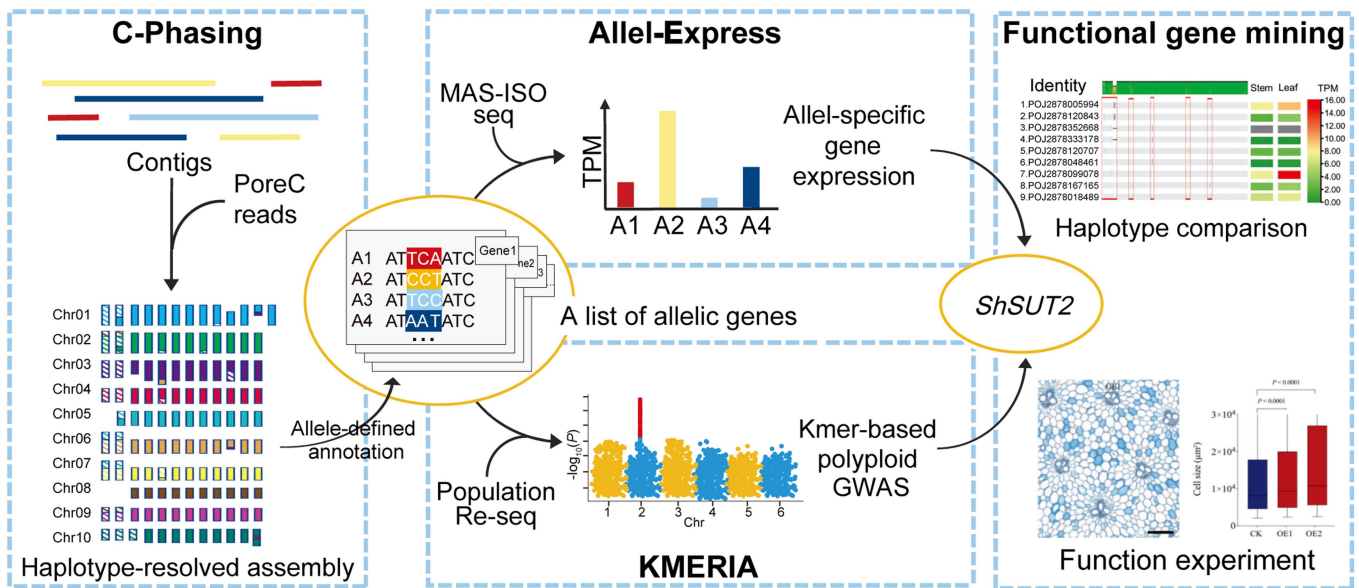
Extended Data Fig. 7 | Validation and Analysis of Parenchyma Cell Counting using Cellpose. (a) Representative original image of a sugarcane stem cross-section used for cell counting. For each group, the microscopic observations were repeated three times independently with similar results. (b) Corresponding classified image, with parenchyma cells identified in white and other cell types in grey. (c) Cellpose segmentation results overlaid on the original image, with green outlines indicating segmented parenchyma cells and red outlines indicating segmented non-parenchyma cells. (d-e) Scatter

plots illustrating the strong correlation between manual and Cellpose-based measurements of cell density (d) and cell size (e), demonstrating the accuracy of the automated method. Statistical significance was determined using a two-sided Pearson correlation test. The r values represent the Pearson correlation coefficients ($n = 10$ independent biological samples). (f-h) Analysis of parenchyma cell density (f) and cell size (g) distributions and their correlation (h) across 265 *Saccharum* hybrid sugarcane cultivars.



Extended Data Fig. 8 | Phenotypic analysis of *ShSUT2*-overexpressed sugarcane lines. (a) Cross-sectional views of the 12th internode from 6-month-old CK (XTT22), OE1 (overexpression line 1), and OE2 (overexpression line 2) plants. Scale bar = 500 μm. (b) Statistical comparison of parenchyma cell size between CK and overexpressed lines was performed using a two-sided Wilcoxon rank-sum test. The lowest 10% and largest 10% outliers were removed from the comparison. Box plots show the median (center line), 25th and 75th

percentiles (box limits), and minimum/maximum values (whiskers); $n = 677$ (CK), 570 (OE1), and 634 (OE2) cells. (c) Relative expression levels (qRT-PCR) of *ShSUT2* and stem sucrose content (HPLC analysis) in CK, OE1, and OE2. Data are presented as mean $\pm$ SD ($n = 3$ independent biological replicates). Statistical significance was determined by two-sided Student's t-test. (d) Verification of *ShSUT2* transgenic lines using a PAT/bar lateral flow assay.



Extended Data Fig. 9 | Integrated Analytical Framework for Polyploid Genomics. The integrated workflow includes three computational tools for comprehensive polyploid genome analysis. *C-Phasing* uses Pore-C sequencing to generate haplotype-resolved assembly with allele-defined gene annotation. *Allele-Express* leverages this assembly to quantify allele-specific expression from MAS-ISO seq data, identifying functionally important haplotypes. *KMERIA* performs K-mer-based GWAS on population re-sequencing data,

with significant associations mapped back to the reference assembly for biological interpretation. The workflow demonstrates functional gene mining, exemplified by the identification and validation of *ShSUT2* through haplotype comparison and functional experiments. This integrated approach creates a synergistic framework where high-quality assembly enables precise functional analysis and meaningful interpretation of population-scale genetic data.

Extended Data Table 1 | Genome sequencing, assembly and annotation of POJ2878

Sequencing	Sugarcane cultivar POJ2878
PacBio HiFi	
Raw data (Gb)	405.42
Sequencing depth (×)	40.6
Average reads length (bp)	15,837.50
Reads N50(bp)	161,77
ONT Ultra-long	
Reads exceeding 100 Kb (Gb)	148.64
Sequencing depth (×)	14.86
Average reads length (bp)	161,940
Reads N50(bp)	166,304
Pore-C	
Raw data (Gb)	517.13
Sequencing depth (×)	51.7
Average reads length (bp)	2,598
Reads N50(bp)	4,150
Haplotype-phased chromosomal-level assembly	
Estimated genome size (Gb) per 2 C	10.00
Assembly size (Gb)	10.37
% of estimated genome size	103.7
Contig N50 (Mb)	15.60
Scaffold N50 (Mb)	84.32
Number of groups	118
% of anchored sequences	97.30
Allele-defined gene annotation	
No. of allele-defined genes in So-subgenome	35,082
No. of allele-defined genes in Ss-subgenome	28,851
No. of homeolog-pairs	28,523
BUSCO completeness (%)	98.6

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
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| ☒ | ☐ | 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
<i>Give P values as exact values whenever suitable.</i> |
| ☒ | ☐ | 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

Pair-end re-sequencing was performed using BGI DNBSEQ-T7 platform;
HIFI reads and Isoform sequences were generated using PacBio Revio platform;
Ultra-long Nanopore sequencing was performed using Promethion platform;
Pore-C sequencing was performed using Oxford Nanopore Technologies;
The parenchyma cell statistics were obtained using our custom code (<https://github.com/Chaiyungun/Automated-Parenchyma-Cell-Analysis-in-Sugarcane-Stems>).

Data analysis

Genome assembly, assessment, and annotation: hifiasm(0.19.1-r559), C-Phasing(v0.2.0. r273), Juicebox (v2.17.00), BUSCO(v5.4.3), Meryl (v1.4.1), Merqury (v1.3), LTR_FINDER_parallel (v1.2), LTR_HARVEST_parallel (v1.1), LTR_retriever (v2.9.6), blastn (v2.9.0), Mosdepth (v0.3.9), quarTeT (v1.2.2), GETA (<https://github.com/chenlianfu/geta>), AUGUSTUS (v.3.3.3), GENEWISE (v.2.4.1), EDTA (2.3.0), RepeatMasker (<http://www.repeatmasker.org>), RepeatModeler2(<https://github.com/Dfam-consortium/RepeatModeler>), DIAMOND (v0.7.10), JCVI (v1.4.18), KMC(v3.2.4), Broccoli (<https://github.com/rderelle/Broccoli>).

Allele-specific expression analysis: ccs tool (<https://github.com/PacificBiosciences/ccs>), skera (<https://github.com/PacificBiosciences/skera>), lima (<https://github.com/PacificBiosciences/barcoding>), isoseq refine (<https://github.com/PacificBiosciences/IsoSeq>), mmseqs (v15.6f452), minimap2 (v2.24-r1122), bedtools (v2.31.1), Allele-Express (<https://github.com/xiaoche-edden/Allele-Express>).

Population analysis: GTX (v2.1.0) (<http://www.gtxlab.com/en/product/cat>), VCF2Dis (v1.52), Admixture (v1.3.0), PopLDdecay (v3.42), VCF2PACCluster (v1.40), IBDseq (r1206), SweeD (v4.0.0), RaiSD (v2.9), VCFtools (v0.1.16), XP-CLR (<https://github.com/hardningj/xpclr>), bcftools (v1.10.2), SnpEff (v5.2c).

GWAS: KMERIA (<https://github.com/Sh1ne111/KMERIA>), CAVIAR (<http://genetics.cs.ucla.edu/caviar/>).

Parenchyma cell statistics: Cellpose, custom code (<https://github.com/Chaiyungungun/Automated-Parenchyma-Cell-Analysis-in-Sugarcane-Stems>).

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 raw sequencing data has been deposited in NCBI database under accession number of PRJNA1303125. The genome assembly and annotation are available at figshare database (<https://figshare.com/s/e79249d6ce11539fa105>) or NCBI database under accession number of PRJNA1312687.

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 applicable

Reporting on race, ethnicity, or other socially relevant groupings Not applicable

Population characteristics Not applicable

Recruitment Not applicable

Ethics oversight Not applicable

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

In this study, a total of 981 sugarcane accessions were collected from 19 countries or regions, representing a broad range of genetic and phenotypic variation. This collection comprised 613 hybrid cultivars, 78 *Saccharum officinarum*, and 290 *Saccharum spontaneum*, constituting the largest known sugarcane population sufficient for comprehensive genetic analysis. To assess the contribution of POJ2878 to sugarcane breeding both within and outside China, we selected 573 cultivated varieties with well-defined geographic origins, including 327 Chinese cultivars (bred in China) and 246 non-Chinese cultivars (bred outside China). Due to the difficulty in cross-sectioning of hard and thick sugarcane stalks, a subset of 265 hybrid cultivars was randomly selected for detailed sugar content analysis, parenchyma cell statistics, and subsequent GWAS analysis. 43 *S. robustum* accessions used for the selective sweep analysis of *S. officinarum* were retrieved from the public database NCBI database under BioProjects PRJNA1228676 and PRJNA1209834.

Data exclusions	No data was excluded from the analyses.
Replication	We confirmed that all attempts at replication were successful.
Randomization	For the field-based sugarcane accessions trials, planting plots were randomized using a complete block design to minimize environmental heterogeneity. For experiments involving stem cross-sectioning and Brix measurement, no experimental group were defined; rather, a representative set of plants at identical developmental stages and physiological ages was selected from the population to ensure phenotypic consistency. Because these experiments involved systematic phenotyping across the population (or genotype) rather than a comparative study between randomized experimental groups, group allocation was not applicable. To control for potential covariates, all selected plants were grown under identical field management practices, and measurements were performed using standardized sampling protocols (e.g., harvesting the same internode position).
Blinding	Blinding was applied during parenchyma cell analysis, Brix measurements, KMERIA-based GWAS, and independent replications of phenotypic. For transgenic experiments, blinding was not applicable because material identification and strict segregation of distinct genetic lines were required. Blinding was not possible for transgenic experiments, as the strict segregation and tracking of distinct genetic lines were required.

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	Anti-PAT/bar (Genesino, China)
Validation	Antibody dilution: Not applicable. We used a commercial lateral flow assay (test strip), not a Western blot. The manufacturer pre-fixes the conjugated antibodies to the strips, making them ready-to-use. The Methods section has been updated to clarify this. Validation statements and experiments can be obtained from the following websites: https://www.genesino.cn/pd.jsp?id=9&fromColId=0#_pp=0_722_2

Dual use research of concern

Policy information about [dual use research of concern](#)

Hazards

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

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

Experiments of concern

Does the work involve any of these experiments of concern:

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

Plants

Seed stocks	The 981 sugarcane accessions were provided by the Yunnan Academy of Agricultural Sciences and are currently stored in the National Sugarcane Germplasm Repository, Kaiyuan City, Yunnan Province (23°30'–23°58' N, 103°04'–103°43' E, altitude: 1051 m).
Novel plant genotypes	Overexpression (OE) lines for both sugarcane (ROC22) and rice (ZH11) were generated using <i>Agrobacterium tumefaciens</i> (EHA105)-mediated transformation of embryogenic calli. For sugarcane, OE lines for ShSUT2, ShTIP1, ShCBL1, and ShATG18a were developed. For rice (ZH11), OE lines for ShTIP1, ShCBL1, and ShATG18a were developed.
Authentication	Verification of the overexpression lines was performed through quantitative RT-PCR, Semi-quantitative RT-PCR, or immunoblotting using Basta antibodies.