nature portfolio

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

Supplementary information

Genetic architecture of sugarcane traits in a polyploid genomics framework

In the format provided by the authors and unedited

Supplementary Note 1: Assessment of POJ2878 Genome Assembly

We assessed the genome assembly quality by investigation completeness, contiguity and accuracy (**Supplementary Table 4**). The completed genome assembly encompasses 10.37 Gb of sequences, with 97.30% of sequences anchored to the 118 chromosomes, representing 103.7% of the estimated total genome size. Evaluation using 1,614 conserved core genes revealed a 99.50% completeness, with 98.60% of genes identified as complete and duplicated. Alignment of MGI-seq reads to the genome assembly showed a mapping rate exceeding 99.93%, with 31.38% of reads uniquely aligned. The assembly demonstrated an improved contiguity with a contig N50 of 15.60 Mb, substantially higher than previously published hybrid sugarcane genomes.

Base-level accuracy was assessed using QV scores (consensus quality value), reached 55.39, corresponding to >99.999% accuracy. Switch errors were identified using ONT-UL reads (see **Methods**), revealing a lower proportion of switched regions than previously published hybrid sugarcane genomes (e.g., 2.02% in POJ2878 vs. 15.78% in ZZ1[5]). Read-depth analysis using 500-kb non-overlapping windows revealed that 98.14% of the genome exhibited uniform coverage by short reads and only 0.82% of regions affected by collapsed errors (**Supplementary Figure 5**). The analysis of chromatin contacts, derived from the alignment of Pore-C reads, reveals stronger intra-subgenome interactions than inter-subgenome interactions (**Figure 1d**). Calculation of Hi-C h-trans errors, which measure the proportion of interactions incorrectly linking non-homologous chromosomes or different haplotypes (indicative of potential misassemblies or chimeric joins), showed that the POJ2878 genome has the lowest h-trans error rate at 1.94% among the examined genomes. These results highlight the superior quality of the POJ2878 genome assembly, with most evaluation metrics meeting or exceeding those of previously published hybrid sugarcane genomes.

Supplementary Note 2: Genome Assembly and Evaluation of *S. officinarum* XZ and *S. spontaneum* 82-114

Genome Sequencing

The genomes of *S. officinarum* XZ and *S. spontaneum* 82-114 were sequenced using PacBio HiFi, Oxford Nanopore Technologies (ONT) ultra-long reads, and Pore-C technologies. Detailed protocols for long DNA extraction, library construction, and sequencing are provided in the Methods section of the main text.

Genome Size Estimation

Genome size and heterozygosity were estimated using K-mer Counter KMC[1-3] (v3.2.4) and JCVI[4] (v1.4.18) software. Based on the 21-mer distribution frequency of PacBio HiFi reads, the estimated genome sizes were 6.98 Gb for *S. officinarum* XZ and 7.21 Gb for *S. spontaneum* 82-114. The average monoploid genome sizes were determined to be 873.22 Mb for *S. officinarum* XZ and 721.36 Mb for *S. spontaneum* 82-114.

Contig Assembly

For *S. officinarum* XZ, a preliminary contig assembly was generated using Hifiasm (v0.19.8-r603) with 265.46 Gb (38x coverage) of PacBio HiFi reads and 84.60 Gb (12× coverage) of ONT ultra-long reads, employing the `--ul` parameter (**Supplementary Table 14**). The initial contig assembly was 4.01 Gb. Redundant sequences were filtered using Khaper[5] (v1.0), resulting in a final monoploid assembly of 879.72 Mb with a contig N50 of 84.80 Mb (**Supplementary Table 15**). The assembly achieved 96.7% completeness based on Benchmarking Universal Single-Copy Orthologs (BUSCO)(v5.4.3) assessment[6,7] (**Supplementary Table 16**).

Similarly, for *S. spontaneum* 82-114, Hifiasm[8-10] (v0.19.8-r603) was used for preliminary contig assembly with 230.84 Gb (32× coverage) of PacBio HiFi reads and 97.9 Gb (13× coverage) of ONT ultra-long reads, using the `--ul` parameter (**Supplementary Table 14**). The final monoploid assembly was 766.92 Mb, with a contig N50 of 94.64 Mb (**Supplementary Table 15**) and a BUSCO completeness of 95.6% (**Supplementary Table 16**).

Monoploid Chromosomal-Scale Assembly

To achieve chromosomal-scale assemblies, Pore-C reads were mapped to the contig assemblies using C-Phasing (v0.2.0. r273). For *S. officinarum* XZ, 533.64 Gb (76x coverage) of Pore-C data were used, while 297.88 Gb (41× coverage) of Pore-C data were used for *S. spontaneum* 82-114. Manual review and correction were performed using Juicebox[11-13](v2.17.00) (**Supplementary Figure 23-24**), with chromosome numbers and orientations adjusted based on the previously published chromosomal-scale assembly of *Erianthus rufipilus*[14]. The final assemblies exhibited an anchoring rate of 99.69% for *S. officinarum* XZ and 92.9% for *S. spontaneum* 82-114 (**Supplementary Table 17**). After removing a chloroplast contig (located on Chr05) and unanchored contigs, the *S. officinarum* XZ assembly was finalized with ten chromosomes, while the *S. spontaneum* 82-114 assembly comprised eight chromosomes. The two monoploid T2T genomes show a high-level synteny with the same chromosome rearrangement events observed in our previously published *S. spontaneum* AP85-441[15] (**Supplementary Figure 25**).

Gap Filling, Centromere, and Telomere Analyses

Gaps in the genome assemblies were filled using Racon[16] (v1.4.20) by aligning ONT raw reads to chromosomes containing gap regions. For *S. officinarum* XZ, a single gap on Chr01 was filled. For *S. spontaneum* 82-114, seven gaps were identified. One gap on Chr03 and two on Chr05 were filled using quarTeT software[17] (v1.2.1). The remaining

gaps (three on Chr06 and one on Chr07) were resolved by separating the chromosomes and applying Racon individually to each.

Centromeres and telomeres were identified and annotated using the `quartet_centrominer.py` and `quartet_TeloExplorer.py` scripts, respectively, from the quarTeT toolkit (v1.2.1) (Lin et al., 2023). All centromeres were identified in both genomes. In *S. officinarum* XZ, telomeres were missing at one end of Chr02, Chr03, Chr04, and Chr05. In *S. spontaneum* 82-114, telomeres were missing at one end of Chr03, Chr05, Chr07, and Chr08, and both ends of Chr06. ONT reads containing the sequences AAACCCT or AGGGTTT were identified and mapped to the genome using Minimap 2[18,19] (v2.24-r1122) to confirm telomere locations (**Supplementary Table 18**).

Supplementary Note 3: Towards accurate and robust quantification of allele-specific gene expression with Allele-Express

1. Motivation

Polyploid species that carry more than one homologous set of chromosomes[20], are widespread among plants and include major crops such as sugarcane, wheat, and potato. Accurate measurement of gene expression in these species is essential for elucidating gene function, understanding genome evolution, and informing crop-improvement efforts[21]. However, the high sequence similarity among allelic and homoeologous copies, together with abundant repetitive elements, complicates the assignment of sequencing reads to specific gene variants[22,23]. Traditional RNA sequencing (RNA-seq) often fails to resolve expression contributions from highly similar allelic copies in polyploid genomes, resulting in systematic biases in allele-specific expression (ASE) quantification[24-26]. Long-read platforms such as PacBio Iso-Seq, and Oxford Nanopore Technologies (ONT) provide alternative approaches, but their relatively low throughput, higher per-read error rates, and platform-specific biases limit their utility for accurate ASE analysis in complex polyploid contexts[27-29]. Computational tools (for example, RSEM[30], Kallisto[31], and IsoQuant[32]) can be challenged by misassignments during alignment, dependence on incomplete SNP datasets, and scalability issues when applied to large, highly redundant assemblies.

To overcome these challenges, we developed Allele-Express, a novel algorithm that leverages MAS-ISO-seq long-read sequencing to perform allele-specific expression (ASE) analysis in haplotype-resolved genome assemblies. By integrating clustering-based strategies with stable point-assisted localization, Allele-Express substantially improves the precision of assigning expression signals to specific alleles while accommodating the genomic complexity of polyploid species. Validation using RNA-seq and PCR experiments across multiple genomes demonstrates its superior performance relative to existing methods.

2. Overview of Allele-Express

The Allele-Express is a comprehensive framework for quantifying allele-specific gene expression, integrating two core modules: **Allele Identification** and **Expression Quantification**.

The **Allele Identification** module is designed to accurately identify allelic genes within haplotype-resolved genome assemblies of both diploid and polyploid organisms (**Extended Data Fig. 4**). Allelic genes, which are alternative forms of a gene found at the same locus on homologous chromosomes, typically exhibit high sequence similarity due to shared evolutionary origins. To distinguish true alleles from other gene family members, we have implemented a multi-faceted strategy. The process begins with a self-alignment of all annotated genes to identify synteny blocks, which are regions of conserved gene order across homologous chromosomes. A minimum of five gene pairs is required to define a synteny block, ensuring that the identified allelic relationships are supported by a conserved genomic context. To account for genes that may have been displaced by structural variations such as rearrangements or inversions, we have incorporated a coordinate-based mapping system. This system utilizes a monoploid reference genome, which can be derived from a closely related species or constructed by removing redundant sequences from the target species' genome. Genes are then classified as allelic based on a large proportion of overlap in their genomic coordinates. To further refine the classification and differentiate alleles from paralogs (genes duplicated within the same genome), the **Allele Identification** module calculates pairwise edit distances among all candidate alleles. It then enforces the constraint that each homologous chromosome can contain only one allelic variant per locus. This rigorous, multi-tiered approach ensures the accurate and reliable classification of allelic relationships.

The **Expression Quantification** module achieves precise measurement of allelic gene expression through an advanced read-to-transcript alignment system (**Extended Data Fig. 4**). The analysis of MAS-ISO-seq reads yields both uniquely and multiply aligned sequences. For reads that align to a single, unique location in the genome, we apply stringent quality control criteria to ensure high-confidence assignments. To resolve the ambiguity of reads that align to multiple locations, we have developed two complementary methods: a read-depth stable point (CSP)-based approach and a read

clustering-based approach. The CSP method identifies a specific point from the beginning of each transcript where the read depth stabilizes, and any reads aligning beyond this point are classified as ambiguous. In the clustering approach, all full-length MAS-ISO-seq reads are subjected to a many-against-many sequence comparison, followed by a greedy incremental clustering algorithm based on overlapping read similarity. This allows for the confident assignment of an ambiguous read to a specific transcript if other reads within the same cluster align uniquely to that transcript. Through this combined strategy, the vast majority of MAS-ISO-seq reads are accurately assigned to their respective transcripts, with only a small fraction being discarded. Finally, the expression level of each transcript is normalized to Transcripts Per Kilobase Million (TPM), providing a robust and standardized measure of allele-specific gene expression.

3. Algorithm Details of Allele-Express

Allele-Express requires high-quality inputs: sequence files for the genome and gene set, annotation files, and optionally coding sequence files for quality checks. These enable accurate grouping of similar genes and further analysis. The algorithm has three main parts—Allele Identification, Quality Control for MAS-ISO-seq reads and expression measurement—which can run separately. Users can start from any step and customize inputs for flexibility.

3.1 Allele Identification

The Allele Identification module employs a multi-stage pipeline to achieve precise allele identification in polyploid genomes. This process consists of five core steps: (1) MCScanX-based synteny analysis, (2) GMAP alignment, (3) BLAST homology search, (4) Transposable element (TE) filtering, and (5) Result refinement.

MCScanX-based Synteny Analysis: We utilize the MCScanX program[33] with default parameters to identify syntenic regions. A BLASTN[34] search is performed on all CDS sequences from the haplotypes, with an E-value threshold of $1x10^{-10}$, retaining the top five hits for each query. A simplified GFF3 file is generated by extracting gene coordinates from the haplotype annotation files. Genes located on the same chromosome

but belonging to different haplotypes are assigned to the same group. Genes within the collinearity blocks identified by MCScanX are collected as candidate allele groups. Tandemly duplicated genes identified by MCScanX are also collected for further evaluation.

GMAP Alignment: A reference genome is either selected or constructed as a monoploid assembly to serve as the GMAP reference. The monoploid assembly can be created by either randomly selecting one haplotype from each chromosome or by using redundancy-removal tools to de-duplicate the contigs, followed by anchoring the non-redundant contigs to the chromosome-scale assembly. All CDS sequences from the haplotypes are then mapped to this reference genome using GMAP[35], with the '-n' parameter set to the ploidy level. The genomic regions of all CDSs are extracted and sorted by their coordinates. We define the alignment overlap ratio (O_{align}) between two regions using the following formula:

$$O_{align} = \frac{L_{gene_1} \cap L_{gene_2}}{L_{gene_1} + L_{gene_2}} \times 200\%$$

, where L_{gene_1} is the length of gene1 and L_{gene_2} is the length of gene2.

If the overlap ratio O_{align} between two regions is ≥ 80%, the corresponding genes are grouped together. Groups that share common genes are subsequently merged. The groups identified by both GMAP and MCScanX are then merged using the same strategy.

BLAST Homology Search: We partition the CDS sequences from all haplotypes into two sets: one containing the genes already assigned to candidate allele groups, and the other containing the remaining genes. We then perform a BLASTN search using the candidate allele genes as the reference and the other genes as the query, with an E-value threshold of 1×10^{-3}. The records with the highest identity for each query gene are collected. If the identity is ≥ 80%, the query gene is added to the corresponding candidate allele group. This step is performed twice by default to ensure comprehensive identification.

TE Overlap Filtering: We identify TE regions using LTR_retriever and then calculate the overlap ratio O_{TE} between each gene and the TE regions using the formula:

$$O_{TE} = \frac{L_{TE} \cap L_{gene}}{L_{gene}}$$

where L_{TE} is the length of TE and L_{gene} is the length of gene.

Genes with an O_{TE} ≥ 30% are excluded from the final allele calls to minimize the impact of transposable elements on allele identification.

Result refinement: For each candidate allele group, genes are first classified into subgroups according to their haplotype of origin. From each subgroup, one representative gene is designated as the allele. Within the same subgroup, genes previously identified as tandem duplicates of the designated allele during the MCScanX synteny analysis step are marked as tandem genes, whereas all remaining genes are designated as paralogs.

3.2 Data Preparation

The Data Preprocessing step is designed to transform raw sequencing data into a high-quality format suitable for downstream analysis. To meet the stringent accuracy requirements of Allele-Express, the raw subreads are first converted into high-quality Kinnex PacBio HiFi reads using the ccs tool (commit v6.4.0; https://github.com/PacificBiosciences/ccs). This conversion is performed using the default parameters (minPasses ≥ 3, minLength = 1000 bp, maxLength = 15,000 bp, minSNR ≥ 4.0, min-rq ≥ 0.99), while allowing users to specify custom parameters when necessary. Following this, the skera tool (v1.2.0; https://github.com/PacificBiosciences/skera) is employed to split the Kinnex PacBio HiFi reads at adapter positions, generating split sequencing reads (S-reads). These S-reads are then aligned to the reference CDS sequence using pbmm2 (v1.13.1; https://github.com/PacificBiosciences/pbmm2) for a comprehensive split evaluation, which includes assessment of CDS coverage and read length distribution. Subsequently, the lima tool (v2.9.0; https://github.com/PacificBiosciences/barcoding) is used to remove the 5' and 3' adapter sequences from the S-reads. Finally, the isoseq3 refine function (commit v4.0.0; https://github.com/PacificBiosciences/isoseq) is applied to remove poly(A) tails and any artificial adapter concatemers, resulting in full-length, non-chimeric (FLNC) sequencing reads. The isoseq3 refine function utilizes the “--require-polya” parameter, which

specifically filters for FLNC reads that contain a poly(A) tail of at least 20 base pairs (bp) in length and then removes these tails.

3.3 Transcript Quantification

The filtered full-length, non-chimeric sequencing reads are aligned to the reference transcriptome using minimap2 (v2.27-r1193)[18]. To ensure accurate read assignment to specific allelic transcripts (a key challenge in polyploid genomes due to high sequence similarity among alleles), the module sorts these alignment results by a sequence of quality metrics in a prioritized order: first by MAPQ (Mapping Quality Score) in descending order (higher values indicate more unique and reliable alignments and minimize ambiguity across multiple matches), followed by edit distance in ascending order (lower values mean fewer base mismatches, insertions, or deletions, reflecting higher sequence similarity between the read and reference transcript), then by AS tag (Alignment Score) in descending order (higher values represent better overall matching, calculated as match rewards minus mismatch penalties), and finally by DE tag (Deletion/Insertion Penalty Score) in ascending order (lower values indicate fewer insertions or deletions, avoiding fragmented, error-prone alignments). The alignment that ranks first after this multi-tiered sorting is selected for downstream analyses, including read assignment in allele-specific expression quantification.

This hierarchical sorting ensures that the highest-quality alignments, characterized by the smallest edit distances and the best alignment scores, are prioritized and defined as unique alignments. For reads that align to multiple locations (multi-alignments), we introduced two novel algorithms—**Reads Clustering** and **CSP Finder**—to optimize the alignment resolution. These algorithms are designed to enhance accuracy and fully leverage the information embedded within the frequent multi-alignments that are characteristic of polyploid genomes.

Reads Clustering. The Reads Clustering algorithm plays a pivotal role in the optimal matching and filtering process, particularly when dealing with reads that have multiple potential alignments, as it effectively addresses the challenge of priority selection. When a sequencing read satisfies the similarity alignment criteria for multiple candidate transcripts simultaneously, the following logic is applied to determine the optimal target transcript.

First, the linear-time clustering algorithm implemented in MMseqs2 (v15.6f452)[36] is used with moderately strict thresholds of 90% sequence identity and 80% query coverage ("easy-linclust --cluster-mode 2 --cov-mode 0 --seq-id-mode 2") to cluster the sequencing reads into groups. The representative sequence for each cluster is defined as the earliest member of that cluster. If the representative sequence has already been assigned to a specific target transcript, the reads within the current cluster are preferentially matched to that same target transcript.

CSP (Coverage Stability Point) Finder. The CSP Finder algorithm is designed to optimize the accuracy of transcript quantification by pinpointing transcript regions with minimal coverage noise. This is achieved through a sliding window approach that analyzes variations in transcript window depth. In this process, the deduplicated reference transcripts are first divided into fixed-size, non-overlapping windows of 10 bp. The number of sequencing reads covering each window (the window depth) is then calculated to produce a window depth file, denoted as D. The CSP Finder takes this window depth file as input, groups the data by transcript, and calculates the window depth fluctuation for each transcript. Regions with minimal window depth variation are then identified and marked as "CSPs". The specific methodology is as follows:

(1) **Grouping of Window Depth Data in File D**

The data in the window depth file *D* are grouped by transcripts. For each transcript, the data are represented as:

$$T_k = \{(s_{k,i}, e_{k,i}, r_{k,i}) | i = 1,2,\dots,n\}$$

, where $s_{k,i}$ and $e_{k,i}$ are the start and end positions of the i-th window respectively. $r_{k,i}$ is the window depth of window i. T_k is the k-th transcript.

(2) **Definition of Sliding Window:** For each transcript T_k, a sliding window $W_{k,j}$ is defined using five consecutive windows:

$$W_{k,j} = r_{k,j}, r_{k,j+1}, r_{k,j+2}, r_{k,j+3}, r_{k,j+4}, \quad j \in 1,2,\dots,n_k - 4$$

, where n_k is the total number of windows for transcript T_k.

(3) **Standard Deviation Calculation:** For each sliding window $W_{k,j}$, the standard

deviation $\sigma(W_{k,j})$ is calculated as:

$$\sigma(W_{k,j}) = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}\left(r_{k,j+i} - \mu(W_{k,j})\right)^2}$$

where $\mu(W_{k,j})$ is the mean of the window depths:

$$\mu(W_{k,j}) = \frac{1}{n}\sum_{i=0}^{n-1} r_{k,j+i}$$

(4) **Identification of CSP.** If the standard deviation $\sigma(W_{k,j})$ is less than the predefined threshold $\tau(\tau = 1.0)$, the end position of the window $e_{k,j}$ is marked as a stability points:

$$s_k = e_{k,j}, \text{ if } \sigma(W_{k,j}) < \tau$$

Once the first window $W_{k,j}$ satisfying the condition （$\sigma(W_{k,j}) < \tau$） is found, the end position of the current window is directly recorded as the stability point, and the loop is terminated.

(5) **No CSP Marking.** If none of the windows for transcript T_k satisfy the stability condition, s_k is marked as None:

$$s_k = \text{None}, \text{ if and only if } \forall j, \sigma(W_{k,j}) \geq \tau$$

Through this mechanism, Allele-Express can effectively filter sequencing reads in low-noise regions, thereby improving the reliability and accuracy of overall quantification analysis results.

Normalization Method. To accurately assess gene expression levels, this study developed a normalization method tailored to the characteristics of sequencing technologies. This section discusses the impact of gene length, sequencing read properties, and sequencing depth on expression quantification. Based on the features of second-generation sequencing data and MAS-ISO-seq data, an appropriate normalization strategy is proposed.

In gene expression quantification, the characteristics of sequencing data largely determine the choice of normalization strategy[37]. Sequencing reads generated from next-generation sequencing (NGS) are typically short fragments (50–300 bp) that cannot

span entire transcripts. As a result, longer transcripts generate more reads, introducing length-dependent bias that inflates their gene expression[38,39]. To address this, the traditional TPM method incorporates transcript length correction, thereby enabling comparisons of gene expression levels within the same sample.

In contrast, MAS-ISO-seq technology produces full-length reads that directly represent complete transcripts. In this context, transcript length and read count are no longer correlated, making traditional length correction unnecessary. However, correction for sequencing depth remains essential to ensure cross-sample comparability. For MAS-ISO-seq data, sequencing depth normalization effectively reduces sample-to-sample variability, thereby improving comparability and enhancing the accuracy of downstream analyses. Based on this consideration, we simplified the normalization workflow for MAS-ISO-seq data by retaining only the sequencing depth correction step and eliminating redundant length normalization steps. This approach not only reduces computational complexity but also better reflects the unique properties of full-length transcriptome sequencing. Specifically, the normalization formula we adopted is:

$$\mathrm{TPM} = \frac{reads\ count}{total\ reads\ count} 10^6$$

4. **Validation**

4.1 Data Simulation

To create a synthetic dataset that closely mimics real sequencing scenarios, we developed a hybrid simulation pipeline integrating PacBio circular consensus sequencing (CCS) and Illumina RNA-seq data. This approach ensures accurate modeling of sequencing errors and transcript expression profiles, providing a robust foundation for evaluating the performance of the Allele-Express algorithm.

PacBio CCS Data Simulation. PacBio CCS read lengths were simulated using the IsoSeqSim tool (https://github.com/yunhaowang/IsoSeqSim), with the POJ2878 genome assembly and annotation as a reference. The tool emulates sequencing errors observed on the Sequel I platform, including substitution (1.731%), deletion (1.090%), and insertion

(2.204%) rates, as well as 5′ and 3′ end truncation effects using predefined probability files (5_end_completeness.PacBio-Sequel.tab and 3_end_completeness.PacBio-Sequel.tab). The simulation outputs include FASTA files containing synthetic read sequences (simulated_ISO-seq_reads.fa) and transcript-level expression quantification in GPD format. These transcript-level data are further processed into gene- and transcript-level count matrices (sim_gene_raw_counts.txt and sim_transcripts_raw_counts.csv), providing comprehensive information for downstream analysis.

Illumina RNA-seq Data Simulation. Illumina RNA-seq data were simulated using the polyester R package, which is designed to generate RNA-seq datasets with differential transcript expression. Transcript expression levels were directly derived from the IsoSeqSim output (sim_transcripts_raw_counts.csv), ensuring consistency between the PacBio and Illumina datasets. Biological replicates were simulated using the generate_reps function, assigning read proportions across three groups (37%, 35%, and 28%) and fixing the random seed (seed = 123) to ensure reproducibility. The final output includes paired-end FASTQ files and raw count matrices (rep_RNA-seq_raw_counts.txt), which align with the simulated PacBio data for integrated analysis.

This hybrid simulation pipeline provides a realistic and reproducible framework for benchmarking allele-specific expression quantification methods, leveraging complementary sequencing technologies to model diverse transcriptomic scenarios.

4.2 RNA-seq Comparative Analysis

We performed a comparative analysis of gene expression quantification accuracy between two computational approaches in the complex polyploid sugarcane genome (*Saccharum hybrid* cultivar POJ2878). Specifically, we compared Kallisto, a widely used RNA-seq quantification algorithm, with Allele-Express, our MAS-ISO-seq–based method (**Supplementary Figure 26**).

Performance of RNA-seq based Kallisto algorithm

Scatter plot analysis revealed substantial limitations of Kallisto in quantifying gene expression in sugarcane. The algorithm achieved a coefficient of determination (R^2) of

0.700, indicating that only ~70% of the variance in simulated ground-truth expression values was explained.

The scatter plot showed considerable dispersion around the regression line, with many genes showing substantial discrepancies between predicted and actual expression levels. This pattern is particularly evident in the mid-to-high expression range (raw counts 10-30). These errors likely stem from Kallisto's pseudoalignment algorithm, which struggles to assign reads correctly in the presence of highly similar homologous sequences characteristic of polyploid genomes.

Performance of MAS-ISO-seq based Allele-Express algorithm

In stark contrast, Allele-Express demonstrates exceptional accuracy in quantifying gene expression within the same polyploid sugarcane genome. The algorithm achieved a remarkably high coefficient of determination (R^2) of 0.953, indicating that approximately 95% of the variance in ground truth expression values is accurately captured by Allele-Express predictions.

The scatter plot reveals a tight clustering of data points along the diagonal regression line, demonstrating consistent accuracy across the entire dynamic range of gene expression levels. This superior performance is evident from low-expression genes (raw counts < 10) through highly expressed genes (raw counts > 30), indicating that Allele-Express maintains quantification precision regardless of transcript abundance.

Comparative Analysis and Implications

The dramatic difference in performance between these two approaches (R^2 = 0.700 vs. 0.953) underscores the critical importance of specialized algorithms for polyploid genome analysis. While Kallisto's pseudoalignment strategy proves effective in diploid genomes, it encounters fundamental limitations when confronted with the high sequence similarity among homologous chromosomes in polyploid species.

The superior performance of Allele-Express can be attributed to several key algorithmic innovations: (1) the utilization of long-read MAS-ISO-seq data that provides unambiguous transcript identification, (2) the implementation of clustering-based read

assignment that leverages sequence similarity patterns, and (3) the Coverage Stability Point (CSP) finder algorithm that identifies high-confidence quantification regions.

These results provide compelling evidence that traditional RNA-seq quantification methods, while adequate for simple genomes, are insufficient for accurate gene expression analysis in complex polyploid systems. The 1.3-fold improvement in R^2 values (0.953 vs. 0.700) represents a substantial advancement in quantification accuracy that has significant implications for downstream biological interpretations and functional analyses in polyploid species.

4.3 Quantitative RT-PCR Validation

The qRT-PCR validation experiments provide compelling evidence for the superior accuracy of Allele-Express compared to conventional RNA-seq quantification methods across different genomic complexities. These results demonstrate the biological relevance and practical utility of our novel algorithm through direct comparison with gold-standard experimental measurements (**Supplementary Figure 27**).

Validation in *Arabidopsis thaliana*

In the diploid model organism *A. thaliana*, both computational approaches (RNA-seq based RSEM algorithm and MAS-ISO-seq based Allele-Express algorithm) showed strong correlations with experimental qRT-PCR measurements, but with notable differences in accuracy. Allele-Express demonstrated exceptional concordance with qRT-PCR quantification, achieving a Pearson correlation coefficient of $r = 0.966$ ($p = 3.17\times10^{-7}$). This remarkably high correlation indicates that Allele-Express captures approximately 93% of the variance in experimental gene expression measurements ($r^2 \approx 0.93$).

In comparison, RSEM, a widely-used RNA-seq quantification tool, showed a lower but still substantial correlation with qRT-PCR data ($r = 0.876$, $p = 1.9\times10^{-4}$). While this correlation is statistically significant and represents reasonable performance for a diploid genome, it falls notably short of the accuracy achieved by Allele-Express.

The scatter plot reveals that both methods generally follow the expected linear relationship with qRT-PCR measurements across the tested dynamic range. However, Allele-Express shows tighter clustering around the regression line, with fewer outliers and

more consistent performance across different expression levels. The confidence intervals (shown in blue and purple shading) demonstrate that Allele-Express maintains higher precision throughout the expression range, particularly for moderately to highly expressed genes.

Validation in hybrid sugarcane

The validation results in the complex polyploid sugarcane genome reveal even more pronounced differences between the two computational approaches, highlighting the critical importance of specialized algorithms for polyploid transcriptomics. Allele-Express maintained robust performance with a correlation coefficient of r = 0.759 ($P = 5.73\times10^{-9}$).

In stark contrast, RSEM showed substantially degraded performance in the sugarcane system, achieving only a moderate correlation of r = 0.444 ($P = 3.25\times10^{-3}$). This represents a dramatic reduction in accuracy compared to its performance in *Arabidopsis*, illustrating the fundamental challenges that traditional RNA-seq quantification methods face when applied to polyploid genomes. The difference in correlation coefficients between the two methods (Δr = 0.315) is particularly striking and represents a 71% improvement in correlation strength.

The scatter plot for sugarcane data reveals distinct patterns that underscore these performance differences. Allele-Express shows a clear linear relationship with qRT-PCR measurements, maintaining reasonable precision across the expression range despite the genomic complexity. The data points cluster relatively tightly around the regression line, with the confidence interval indicating consistent performance.

RSEM's performance in sugarcane shows considerably more scatter, with numerous data points deviating substantially from the expected linear relationship. The wider confidence interval and flatter regression slope suggest that RSEM struggles to capture the true dynamic range of gene expression in this polyploid system. This poor performance likely stems from the algorithm's inability to accurately assign reads among highly similar homologous sequences, leading to systematic biases in expression quantification.

These experimental validation results, combined with the computational benchmarking studies, provide compelling evidence that Allele-Express represents a

significant advancement in gene expression quantification, particularly for complex polyploid genomes where traditional methods show substantial limitations.

Supplementary Note 4: An integrated genomics framework for polyploid crops

This study, together with our companion studies[40,41], established an integrated analytical framework to tackle the complexities of polyploid crop genomes by introducing three complementary tools (**Extended Data Fig. 9**). These tools enable precise assembly, analysis, and interpretation of polyploid genomes, facilitating deeper biological insights and breeding applications.

The *C-Phasing* package[40] leverages Pore-C sequencing data to generate haplotype-resolved assemblies at chromosomal scale. This tool addresses the challenges in polyploid genomics—accurately separating homologous and homoeologous sequences. The resulting POJ2878 assembly comprises 118 individually phased chromosomes across 10 homoeologous groups, creating an essential sugarcane reference genome for all downstream analyses.

KMERIA[41] performs K-mer-based GWAS, which avoids the errors inherent in complex polyploid genotyping. This approach further leverages a high-quality reference genome to facilitate the biological interpretation of significant K-mer associations. While previous K-mer-based GWAS methods, such as those developed by Voichek [42] et al. and He et al., [43] laid important groundwork for diploid species like maize, their models are not readily applicable to polyploid genomes. *KMERIA* extends these approaches through allele-dosage modeling and integrates a linear mixed model (LMM) to correct for population structure and kinship, improving both robustness and precision. Implemented in optimized C/C++ with algorithmic refinements, *KMERIA* also enables scalable and statistically sound association analyses in complex polyploid systems.

Allele-Express directly depends on the haplotype-resolved assembly and allele-defined annotation for accurate allele-specific expression quantification for these genes identified by *KMERIA* or selection. Without precisely phased chromosomes,

distinguishing transcripts from different alleles in such a complex polyploid would be impossible.

For example, K-mers associated with parenchyma cell traits were mapped to the POJ2878 assembly, leading to the identification of candidate genes such as *ShSUT2*. Using the phased genome, we identified allelic variants across different haplotypes, including one specific allele showing high expression in leaf tissue. This example demonstrates that the integrated genomics framework provides a powerful platform for targeting key alleles for further functional validation or as breeding targets.

References

1 Kokot, M., Długosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759-2761 (2017). https://doi.org/10.1093/bioinformatics/btx304

2 Deorowicz, S., Kokot, M., Grabowski, S. & Debudaj-Grabysz, A. KMC 2: fast and resource-frugal k-mer counting. *Bioinformatics* **31**, 1569-1576 (2015). https://doi.org/10.1093/bioinformatics/btv022

3 Deorowicz, S., Debudaj-Grabysz, A. & Grabowski, S. Disk-based k-mer counting on a PC. *BMC Bioinformatics* **14**, 160 (2013). https://doi.org/10.1186/1471-2105-14-160

4 Tang, H. *et al.* JCVI: A versatile toolkit for comparative genomics analysis. *iMeta* **3** (2024).

5 Zhang, X. *et al.* Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant Camellia sinensis. *Nature Genetics* **53**, 1250-1259 (2021). https://doi.org/10.1038/s41588-021-00895-y

6 Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* **38**, 4647-4654 (2021). https://doi.org/10.1093/molbev/msab199

7 Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: Assessing Genomic Data Quality and Beyond. *Current Protocols* **1**, e323 (2021). https://doi.org/https://doi.org/10.1002/cpz1.323

8 Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170-175 (2021). https://doi.org/10.1038/s41592-020-01056-5

9 Cheng, H., Asri, M., Lucas, J., Koren, S. & Li, H. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. *Nature Methods* **21**, 967-970 (2024). https://doi.org/10.1038/s41592-024-02269-8

10 Cheng, H. *et al.* Haplotype-resolved assembly of diploid genomes without parental data. *Nature Biotechnology* **40**, 1332-1335 (2022). https://doi.org/10.1038/s41587-022-01261-x

11 Dudchenko, O. *et al.* The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. *bioRxiv*, 254797 (2018). https://doi.org/10.1101/254797
12 Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems* **3**, 99-101 (2016). https://doi.org/https://doi.org/10.1016/j.cels.2015.07.012
13 Rao, Suhas S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665-1680 (2014). https://doi.org/https://doi.org/10.1016/j.cell.2014.11.021
14 Wang, T. *et al.* A complete gap-free diploid genome in *Saccharum* complex and the genomic footprints of evolution in the highly polyploid *Saccharum* genus. *Nature Plants* **9**, 554-571 (2023). https://doi.org/10.1038/s41477-023-01378-0
15 Zhang, J. *et al.* Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. *Nature Genetics* **50**, 1565-1573 (2018). https://doi.org/10.1038/s41588-018-0237-2
16 Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *bioRxiv* (2016). https://doi.org/10.1101/068122
17 Lin, Y. *et al.* quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Horticulture Research* **10**, uhad127 (2023). https://doi.org/10.1093/hr/uhad127
18 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018). https://doi.org/10.1093/bioinformatics/bty191
19 Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572-4574 (2021). https://doi.org/10.1093/bioinformatics/btab705
20 Comai, L. The advantages and disadvantages of being polyploid. *Nature Reviews Genetics* **6**, 836-846 (2005). https://doi.org/10.1038/nrg1711
21 Ramírez-González, R. H. *et al.* The transcriptional landscape of polyploid wheat. *Science* **361**, eaar6089 (2018). https://doi.org/doi:10.1126/science.aar6089
22 Kong, W., Wang, Y., Zhang, S., Yu, J. & Zhang, X. Recent Advances in Assembly of Complex Plant Genomes. *Genomics, Proteomics & Bioinformatics* **21**, 427-439 (2023). https://doi.org/10.1016/j.gpb.2023.04.004
23 Cheng, L. *et al.* Genome analyses and breeding of polyploid crops. *Nature Plants* **11**, 1714-1728 (2025). https://doi.org/10.1038/s41477-025-02088-5
24 Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nature Reviews Genetics* **20**, 631-656 (2019). https://doi.org/10.1038/s41576-019-0150-2
25 Turro, E. *et al.* Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biology* **12**, R13 (2011). https://doi.org/10.1186/gb-2011-12-2-r13
26 Oshlack, A., Robinson, M. D. & Young, M. D. From RNA-seq reads to differential expression results. *Genome Biology* **11**, 220 (2010). https://doi.org/10.1186/gb-2010-11-12-220
27 Method of the Year 2022: long-read sequencing. *Nature Methods* **20**, 1-1 (2023). https://doi.org/10.1038/s41592-022-01759-x

28 Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biology* **21**, 30 (2020). https://doi.org/10.1186/s13059-020-1935-5

29 Monzó, C., Liu, T. & Conesa, A. Transcriptomics in the era of long-read sequencing. *Nature Reviews Genetics* **26**, 681-701 (2025). https://doi.org/10.1038/s41576-025-00828-z

30 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011). https://doi.org/10.1186/1471-2105-12-323

31 Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34**, 525-527 (2016). https://doi.org/10.1038/nbt.3519

32 Prjibelski, A. D. *et al.* Accurate isoform discovery with IsoQuant using long reads. *Nature Biotechnology* **41**, 915-918 (2023). https://doi.org/10.1038/s41587-022-01565-y

33 Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* **40**, e49-e49 (2012). https://doi.org/10.1093/nar/gkr1293

34 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410 (1990). https://doi.org/https://doi.org/10.1016/S0022-2836(05)80360-2

35 Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-1875 (2005). https://doi.org/10.1093/bioinformatics/bti310

36 Kallenborn, F. *et al.* GPU-accelerated homology search with MMseqs2. *bioRxiv*, 2024.2011.2013.623350 (2024). https://doi.org/10.1101/2024.11.13.623350

37 Dillies, M.-A. *et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics* **14**, 671-683 (2012). https://doi.org/10.1093/bib/bbs046

38 Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences* **131**, 281-285 (2012). https://doi.org/10.1007/s12064-012-0162-3

39 Abrams, Z. B., Johnson, T. S., Huang, K., Payne, P. R. O. & Coombes, K. A protocol to evaluate RNA sequencing normalization methods. *BMC Bioinformatics* **20**, 679 (2019). https://doi.org/10.1186/s12859-019-3247-x

40. Wang, Y. *et al.* Enhanced Pore-C with C-Phasing Enables Chromosomal-Scale, Haplotype-Resolved Assembly of Ultra-Complex Genomes. *PREPRINT (Version 1) available at Research Square* https://doi.org/10.21203/rs.3.rs-7343323/v1 (2025)

41. Chen, S. *et al.* A k-mer-based GWAS approach empowering gene mining in polyploids. *PREPRINT (Version 1) available at Research Square* https://doi.org/10.21203/rs.3.rs-7347406/v1 (2025)

42. Voichek, Y. & Weigel, D. Identifying genetic variants underlying phenotypic variation in plants without complete genomes. *Nature Genetics* **52**, 534–540 (2020).

43. He, C. *et al.* Trait association and prediction through integrative *k* -mer analysis. *The Plant Journal* **120**, 833–850 (2024).

Supplementary Table 1. Statistics of HiFi and ONT-UL sequencing reads

Items	HiFi	ONT-UL	MGI-seq
Total reads	25,599,002	917,888	4,523,448,704
Total bases (Gb)	405.42	148.64	678.52
Mean length (bp)	15,837.50	161,940	149.25
Max length (bp)	50,344	1,228,391	150
Q10	12,427	119,962	150
Q20	14,840	147,502	150
Q30	18,298	188,572	150
N50 (bp)	16,177	166,304	150
GC(%)	44.54	44.67	44.93

Supplementary Table 2. Statistics of Pore-C sequencing reads

Items	Pore-C	HiC
Total reads (million)	199.02	8,383.89
Total bases (Gb)	517.13	1,254.94
Mean length (bp)	2,598	149.7
Max length (bp)	840,982	150
Q10	838	150
Q20	1,710	150
Q30	3,375	150
N50	4,150	150
GC content (%)	44.37	44.21
Uniquely aligned reads	147.25	2,489.18
Uniquely aligned reads (%)	77.99	39.81
Valid pairs of interaction (%)	36.44	29.69
Order = 2 (%)	21.29	29.69
Order ≥ 3 (%)	15.15	0

Supplementary Table 3. Statistics of contig-level assembly

Items	POJ2878
No. of contigs	14,029
Total length (Gb)	10.37
Maximum length (Mb)	113.86
N50 (Mb)	15.60
N90 (Kb)	370.60
Average length (Kb)	739.37

Supplementary Table 4. Assessment and comparison of genome assemblies in hybrid sugarcane

Species	POJ2878	R570	ZZ01	XTT22
Contiguity				
Contig N50 (Mb)	**15.60**	8.20	0.13	0.51
Gap number per Gb	794	**591**	11937	4570
Completeness				
Assembly size (Gb)	10.37	8.77	10.38	9.28
% of estimated genome size	**103.70**	87.70	115.30	-
BUSCO completeness (%)	99.50	98.90	**99.70**	99.60
K-mer completeness (%)	**96.78**	93.66	96.02	96.43
Mapping rate (NGS; %)	**99.93**	99.63	99.73	99.92
Raw LAI	20.88	**22.53**	18.39	20.38
Anchor rate (%)	**97.30**	70.52	87.25	95.25
Accuracy				
k-mer QV	**55.39**	45.81	49.76	36.40
Hi-C h-trans error (%)	**1.94**	-	3.43	3.51
Switch error (%)	**2.02**	-	15.78	-
Collapsed error (%)	**0.82**	12.64	5.82	8.27

- Missing assessment due to unavailable data.

Supplementary Table 5. The genetic component of POJ2878

ChrID	subgenome	Length (Mb)	So-subgenome component		Ss-subgenome component		Homeologous exchange (HE)		
			So sizes (Mb)	So perct (%)	Ss sizes (Mb)	Ss perct (%)	HE (Mb)	sizes	HE perct (%)
Chr01g1	Ss	102.71			102.60	99.89		0.11	0.11
Chr01g2	Rec	106.36	40.39	37.97	65.97	62.03			
Chr01g3	So	131.71	131.71	100.00					
Chr01g4	So	126.75	126.75	100.00					
Chr01g5	So	125.37	125.13	99.81				0.24	0.19
Chr01g6	So	125.30	125.06	99.81				0.24	0.19
Chr01g7	So	122.69	121.74	99.23				0.95	0.77
Chr01g8	So	123.79	123.31	99.61				0.48	0.39
Chr01g9	Rec	120.65	104.87	86.92	15.78	13.08			
Chr01g10	So	82.45	82.34	99.87				0.11	0.13
Chr01g11	So	117.26	117.15	99.91				0.11	0.09
Chr01g12	So	70.81	70.44	99.48				0.37	0.52
Chr01g13	So	131.30	131.30	100.00					
Chr02g1	Ss	113.86			113.86	100.00			
Chr02g2	Rec	118.41	12.76	10.78	105.65	89.22			
Chr02g3	So	109.05	109.05	100.00					
Chr02g4	So	110.57	110.57	100.00					
Chr02g5	Rec	104.03	86.20	82.86	17.83	17.14			
Chr02g6	So	105.61	105.26	99.67				0.35	0.33
Chr02g7	So	105.69	105.58	99.90				0.11	0.10
Chr02g8	Rec	98.41	76.72	77.96	21.69	22.04			
Chr02g9	So	97.65	97.50	99.85				0.15	0.15
Chr02g10	So	104.99	104.71	99.73				0.28	0.27
Chr02g11	So	103.82	103.69	99.87				0.13	0.13
Chr02g12	So	109.93	109.39	99.51				0.54	0.49
Chr03g1	Ss	80.03			79.92	99.86		0.11	0.14
Chr03g2	Ss	87.20			87.20	100.00			
Chr03g3	So	76.27	76.27	100.00					
Chr03g4	So	102.87	102.72	99.85				0.15	0.15
Chr03g5	So	116.54	116.37	99.85				0.17	0.15
Chr03g6	So	101.88	101.88	100.00					
Chr03g7	So	102.68	102.68	100.00					
Chr03g8	So	101.29	100.97	99.68				0.32	0.32
Chr03g9	So	93.85	93.46	99.58				0.39	0.42
Chr03g10	Rec	91.94	34.29	37.30	57.65	62.70			

Chr03g11	So	101.90	101.90	100.00				
Chr03g12	So	105.17	105.17	100.00				
Chr04g1	Ss	72.72			72.72	100.00		
Chr04g2	Ss	73.80			73.80	100.00		
Chr04g3	So	88.34	88.19	99.83			0.15	0.17
Chr04g4	So	90.70	90.70	100.00				
Chr04g5	Rec	86.64	17.87	20.63	68.77	79.37		
Chr04g6	So	94.89	94.68	99.78			0.21	0.22
Chr04g7	So	90.77	90.64	99.86			0.13	0.14
Chr04g8	So	88.51	88.27	99.73			0.24	0.27
Chr04g9	So	88.75	88.34	99.54			0.41	0.46
Chr04g10	So	85.77	85.77	100.00				
Chr04g11	So	91.13	91.13	100.00				
Chr04g12	So	90.29	90.29	100.00				
Chr05g1	Rec	81.79	32.54	39.78	49.25	60.22		
Chr05g2	So	81.00	81.00	100.00				
Chr05g3	So	90.09	90.09	100.00				
Chr05g4	So	88.87	88.87	100.00				
Chr05g5	So	84.08	83.97	99.87			0.11	0.13
Chr05g6	So	82.56	82.15	99.50			0.41	0.50
Chr05g7	So	85.71	85.71	100.00				
Chr05g8	So	78.46	78.46	100.00				
Chr05g9	So	86.71	86.71	100.00				
Chr05g10	So	79.23	78.71	99.34			0.52	0.66
Chr05g11	So	88.41	88.41	100.00				
Chr06g1	Ss	102.07			102.07	100.00		
Chr06g2	Ss	100.26			100.26	100.00		
Chr06g3	So	77.83	77.57	99.67			0.26	0.33
Chr06g4	So	74.64	74.64	100.00				
Chr06g5	So	75.30	75.30	100.00				
Chr06g6	So	71.24	71.13	99.85			0.11	0.15
Chr06g7	So	69.65	69.65	100.00				
Chr06g8	So	74.61	74.27	99.54			0.34	0.46
Chr06g9	So	73.80	73.80	100.00				
Chr06g10	So	52.48	52.06	99.20			0.42	0.80
Chr06g11	So	68.05	67.79	99.62			0.26	0.38
Chr06g12	So	70.53	70.53	100.00				
Chr07g1	Ss	92.00			92.00	100.00		
Chr07g2	Ss	92.09			92.09	100.00		
Chr07g3	So	68.41	68.30	99.84			0.11	0.16

Chr07g4	So	71.99	71.99	100.00				
Chr07g5	Rec	68.86	10.62	15.42	58.24	84.58		
Chr07g6	So	65.12	64.93	99.71			0.19	0.29
Chr07g7	So	68.95	68.95	100.00				
Chr07g8	So	69.29	69.29	100.00				
Chr07g9	So	70.43	70.30	99.82			0.13	0.18
Chr07g10	So	72.45	72.45	100.00				
Chr07g11	So	66.86	66.62	99.64			0.24	0.36
Chr07g12	So	66.81	66.70	99.84			0.11	0.16
Chr08g1	So	60.69	60.69	100.00				
Chr08g2	So	62.82	62.36	99.27			0.46	0.73
Chr08g3	So	62.27	62.10	99.73			0.17	0.27
Chr08g4	So	62.36	62.25	99.82			0.11	0.18
Chr08g5	So	62.63	62.63	100.00				
Chr08g6	So	59.69	59.69	100.00				
Chr08g7	So	62.12	62.12	100.00				
Chr08g8	So	62.58	62.58	100.00				
Chr08g9	Rec	63.73	54.10	84.89	9.63	15.11		
Chr08g10	So	60.50	60.26	99.60			0.24	0.40
Chr09g1	Ss	80.56			80.56	100.00		
Chr09g2	Ss	82.08			82.08	100.00		
Chr09g3	So	73.59	73.44	99.80			0.15	0.20
Chr09g4	So	76.01	76.01	100.00				
Chr09g5	So	77.10	77.10	100.00				
Chr09g6	So	70.92	70.77	99.79			0.15	0.21
Chr09g7	So	72.39	72.26	99.82			0.13	0.18
Chr09g8	So	74.12	74.12	100.00				
Chr09g9	So	71.89	71.76	99.82			0.13	0.18
Chr09g10	So	70.57	70.57	100.00				
Chr09g11	So	65.37	65.37	100.00				
Chr09g12	So	62.77	62.62	99.76			0.15	0.24
Chr10g1	Ss	67.58			67.58	100.00		
Chr10g2	Ss	62.19			62.19	100.00		
Chr10g3	Ss	65.12			65.12	100.00		
Chr10g4	So	76.89	76.43	99.40			0.46	0.60
Chr10g5	So	80.97	80.97	100.00				
Chr10g6	So	69.92	69.60	99.54			0.32	0.46
Chr10g7	So	77.96	77.72	99.69			0.24	0.31
Chr10g8	So	80.75	80.08	99.17			0.67	0.83
Chr10g9	So	79.15	78.96	99.76			0.19	0.24

Chr10g10	So	76.02	76.02	100.00				
Chr10g11	So	80.86	80.86	100.00				
Chr10g12	So	95.22	95.22	100.00				
	Total	10093.12	8335.38	82.58	1744.51	17.28	13.23	0.13

Supplementary Table 6. Identification and classification of centromeres

Chrom	Position		Length(Mb)	Type
	Physical position (bp)	Relative position (%)*		
Chr01g1	51178859-52524501	50.48	1.35	metacentric
Chr01g2	103427129-106334709	98.61	2.91	telocentric
Chr01g3	65378273-68354831	50.77	2.98	metacentric
Chr01g4	124060525-126751239	98.94	2.69	telocentric
Chr01g5	122238855-125321232	98.73	3.08	telocentric
Chr01g6	59016184-61823743	48.22	2.81	metacentric
Chr01g7	120626840-121795093	98.80	1.17	acrocentric
Chr01g8	59705435-63257680	49.67	3.55	metacentric
Chr01g9	118597630-120635351	99.14	2.04	telocentric
Chr01g10	80042185-81694542	98.08	1.65	acrocentric
Chr01g11	53278951-56965314	47.01	3.69	metacentric
Chr01g12	41373443-43377968	59.84	2.00	submetacentric
Chr01g13	127894081-131283761	98.70	3.39	telocentric
Chr02g1	61974840-63452934	55.08	1.48	metacentric
Chr02g2	2504116-3729332	2.63	1.23	acrocentric
Chr02g3	1-3295821	1.51	3.30	telocentric
Chr02g4	1-1777828	0.80	1.78	telocentric
Chr02g5	1-1616071	0.78	1.62	telocentric
Chr02g6	44290067-45195880	42.37	0.91	metacentric
Chr02g7	42309407-45995293	41.77	3.69	metacentric
Chr02g8	40843209-41211666	41.69	0.37	metacentric
Chr02g9	1-1957511	1.00	1.96	telocentric
Chr02g10	41819772-44272105	41.00	2.45	submetacentric
Chr02g11	39310719-42638919	39.47	3.33	submetacentric
Chr02g12	42721855-45539598	40.14	2.82	submetacentric
Chr03g1	31645383-32607724	40.14	0.96	submetacentric
Chr03g2	35216795-38747124	42.41	3.53	metacentric
Chr03g3	39221909-41825004	53.13	2.60	metacentric
Chr03g4	44265804-45669031	43.71	1.40	metacentric
Chr03g5	40688172-43432679	36.09	2.74	submetacentric
Chr03g6	41430350-43655670	41.76	2.23	metacentric
Chr03g7	43573741-44638142	42.95	1.06	metacentric
Chr03g8	42654886-43965751	42.76	1.31	metacentric
Chr03g9	33305080-36280001	37.07	2.97	submetacentric
Chr03g10	42626823-44503002	47.38	1.88	metacentric
Chr03g11	1-856754	0.42	0.86	telocentric
Chr03g12	42582491-45613164	41.93	3.03	metacentric
Chr04g1	28211607-32061424	41.44	3.85	submetacentric
Chr04g2	28605695-32698868	41.53	4.09	submetacentric

Chr04g3	35402824-38543737	41.85	3.14	metacentric
Chr04g4	35777783-38203839	40.79	2.43	submetacentric
Chr04g5	38118092-41664066	46.04	3.55	metacentric
Chr04g6	39265093-43431014	43.58	4.17	metacentric
Chr04g7	35042607-39176667	40.88	4.13	submetacentric
Chr04g8	33772564-36703198	39.81	2.93	submetacentric
Chr04g9	34962653-38138638	41.19	3.18	submetacentric
Chr04g10	32842837-36289846	40.30	3.45	submetacentric
Chr04g11	36958365-40469982	42.48	3.51	metacentric
Chr04g12	35675806-38495001	41.07	2.82	submetacentric
Chr05g1	36408638-40205430	46.84	3.80	metacentric
Chr05g2	37001028-37937403	46.26	0.94	metacentric
Chr05g3	40899777-41928380	45.97	1.03	metacentric
Chr05g4	38910560-41372654	45.17	2.46	metacentric
Chr05g5	36333546-38349129	44.41	2.02	metacentric
Chr05g6	37504107-39453109	46.61	1.95	metacentric
Chr05g7	38898282-40718861	46.44	1.82	metacentric
Chr05g8	30435350-31413977	39.41	0.98	submetacentric
Chr05g9	36333665-36623946	42.07	0.29	metacentric
Chr05g10	27541844-28350302	35.27	0.81	submetacentric
Chr05g11	35085416-35874056	40.13	0.79	submetacentric
Chr06g1	51735508-54873950	52.22	3.14	metacentric
Chr06g2	53098642-55271293	54.04	2.17	metacentric
Chr06g3	23076895-25171777	31.00	2.09	submetacentric
Chr06g4	19278348-21402908	27.25	2.12	submetacentric
Chr06g5	20629231-21102201	27.71	0.47	submetacentric
Chr06g6	19183889-19608799	27.23	0.42	submetacentric
Chr06g7	21351251-21533228	30.79	0.18	submetacentric
Chr06g8	22841682-23021279	30.74	0.18	submetacentric
Chr06g9	18588324-18785880	25.32	0.20	submetacentric
Chr06g10	17580306-19523593	35.35	1.94	submetacentric
Chr06g11	17368194-17556366	25.66	0.19	submetacentric
Chr06g12	18726136-19092659	26.81	0.37	submetacentric
Chr07g1	88979723-90490281	97.54	1.51	acrocentric
Chr07g2	62355067-63919460	68.56	1.56	submetacentric
Chr07g3	1-3552825	2.60	3.55	telocentric
Chr07g4	37063659-37988474	52.13	0.92	metacentric
Chr07g5	30802262-38683937	50.45	7.88	metacentric
Chr07g6	26558962-30275884	43.64	3.72	metacentric
Chr07g7	33219778-33825612	48.62	0.61	metacentric
Chr07g8	1-200525	0.14	0.20	telocentric
Chr07g9	34366856-38434136	51.68	4.07	metacentric
Chr07g10	1-2164443	1.49	2.16	telocentric

Chr07g11	30541393-31614624	46.48	1.07	metacentric
Chr07g12	34067265-34772664	51.52	0.71	metacentric
Chr08g1	27240191-29735121	46.94	2.49	metacentric
Chr08g2	27857992-30712712	46.62	2.85	metacentric
Chr08g3	28377882-30957458	47.64	2.58	metacentric
Chr08g4	27490737-29722779	45.87	2.23	metacentric
Chr08g5	28889302-31465015	48.18	2.58	metacentric
Chr08g6	27460082-29887066	48.04	2.43	metacentric
Chr08g7	27325488-29881541	46.04	2.56	metacentric
Chr08g8	27501384-30446670	46.30	2.95	metacentric
Chr08g9	28072438-32265745	47.34	4.19	metacentric
Chr08g10	26813771-29320657	46.39	2.51	metacentric
Chr09g1	35698939-37203850	45.24	1.50	metacentric
Chr09g2	3997525-5295659	5.66	1.30	acrocentric
Chr09g3	1-2274939	1.55	2.27	telocentric
Chr09g4	23233464-24634532	31.49	1.40	submetacentric
Chr09g5	33819860-37432229	46.21	3.61	metacentric
Chr09g6	31287731-32326317	44.85	1.04	metacentric
Chr09g7	21213172-21954950	29.82	0.74	submetacentric
Chr09g8	31847383-32899441	43.68	1.05	metacentric
Chr09g9	31036824-31976505	43.83	0.94	metacentric
Chr09g10	30875803-32035640	44.57	1.16	metacentric
Chr09g11	24887802-27752039	40.26	2.86	submetacentric
Chr09g12	28256779-31688069	47.75	3.43	metacentric
Chr10g1	63456206-65656240	95.52	2.20	acrocentric
Chr10g2	30322619-31919697	50.04	1.60	metacentric
Chr10g3	32290130-34448367	51.24	2.16	metacentric
Chr10g4	37544794-40538358	50.77	2.99	metacentric
Chr10g5	38204685-41041734	48.93	2.84	metacentric
Chr10g6	36836711-41094826	55.73	4.26	metacentric
Chr10g7	36984453-37819696	47.97	0.84	metacentric
Chr10g8	37235455-40540313	48.16	3.30	metacentric
Chr10g9	37516466-40467219	49.26	2.95	metacentric
Chr10g10	37350589-40038340	50.90	2.69	metacentric
Chr10g11	41039088-43701794	52.40	2.66	metacentric
Chr10g12	55259680-57952262	59.45	2.69	submetacentric

* Relative position is calculated as a percentage from the 5' start of the chromosome.

Supplementary Table 7. Statistics of Transposable Elements in POJ2878 and two progenitor monoploid genomes (S*. officinarum* XZ and *S. spontaneum* 82-114)

	POJ_So-subgenome		POJ_Ss-subgenome		SoXZ monoploid genome		Ss82-114 monoploid genome	
	Length (Mb)	% of genome	Length (Mb)	% of genome	Length (Mb)	% of genome	Length (Mb)	% of genome
Total Repeat Fraction	5,855.56	70.13	1,326.83	65.68	623.05	71.03	476.81	66.13
Class I: Retroelement	4,935.39	59.11	1,052.55	52.11	434.35	49.52	298.59	41.41
LTR Retrotransposon	4,930.31	59.05	1,051.15	52.04	433.96	49.48	298.21	41.36
Copia	1383.23	16.57	296.87	14.70	138.48	15.79	99.71	13.83
Gypsy	2167.73	25.96	448.79	22.22	210.56	24.01	146.95	20.38
Other	1379.35	16.52	305.49	15.12	84.92	9.68	51.54	7.15
LINE	5.08	0.06	1.39	0.07	0.39	0.04	0.39	0.05
Class II: DNA transposon	879.32	10.53	262.26	12.98	166.69	19.00	166.27	23.06
TIR	678.62	8.13	197.30	9.77	98.45	11.22	86.36	11.98
CMD[DTC]	199.10	2.38	63.98	3.17	27.61	3.15	20.72	2.87

Mutator	177.65	2.13	49.69	2.46	31.26	3.56	29.05	4.03
PIF/Harbinger	117.65	1.41	30.16	1.49	10.53	1.20	8.40	1.17
Tc1/Mariner	122.83	1.47	37.85	1.87	18.46	2.10	19.73	2.74
hAT	61.39	0.74	15.62	0.77	10.60	1.21	8.46	1.17
Helitron	200.70	2.40	64.95	3.22	68.25	7.78	79.91	11.08
Unclassified repeats	40.85	0.49	12.03	0.60	22.00	2.51	11.95	1.66

Supplementary Table 8. BUSCO assessment of gene annotation

Items	Number of genes	Percentage (%)
Complete BUSCOs (C)	1590	98.6
Complete and single-copy BUSCOs (S)	9	0.6
Complete and duplicated BUSCOs (D)	1581	98.0
Fragmented BUSCOs (F)	6	0.4
Missing BUSCOs (M)	18	1.0
Total BUSCO groups searched	1614	100

Supplementary Table 9. Allele-defined annotation of the two subgenomes

ChrGroup	So subgenome			Ss subgenome			No. of homeolog-paris
	No. of genes with defined alleles[1]	No. of haplotype-specific genes[2]	No. of paralogs[3]	No. of genes with defined alleles	No. of haplotype-specific genes	No. of paralogs	
Chr01	5830	797	2076	4825	335	1047	4,795
Chr02	4782	625	2681	4025	433	1458	3,965
Chr03	4648	616	2344	4074	313	1119	4,026
Chr04	3860	471	2182	3438	318	993	3,392
Chr05	2631	652	3063	1652	99	2004	1,645
Chr06	3030	429	1524	2575	304	561	2,547
Chr07	2701	431	1955	2350	356	1154	2,296
Chr08	1910	385	1409	994	179	20	994
Chr09	2703	369	906	2274	259	510	2,256
Chr10	2987	415	1229	2644	305	815	2,607
Total	35,082	5,190	19,369	28,851	2,722	9,681	28,523

1. Genes with defined alleles are characterized as having more than one allele or homoeologous copies across the genomes.
2. Haplotype-specific genes are defined as those that are present only in a single haplotype and absent in other regions of the genome.
3. Paralogs include both tandemly duplicated and distantly duplicated genes that exhibit substantial similarity to allelic genes.

Supplementary Table 10. Statistics of full-length transcripts in leaf and stem tissues

	Leaf tissue			Stem tissue		
	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3
No. Of full-length transcripts (million)	21.65	21.42	20.81	13.64	27.54	15.05
Data Size (Gb)	30.86	31.05	30.33	19.18	41.49	22.82
Average length (bp)	1425	1449	1457	1406	1506	1515
Max length (bp)	8761	9519	9231	9397	9377	8940

Supplementary Table 12. Association between copy numbers of specific K-mers of the SUS2 favorable haplotype and sucrose content at 270 days post-planting

Copy number	Individual number	Log_2(Sucrose content) (mean)	*P-value*
0	8	7.53	-
1	43	7.89	0.047
2	61	7.82	0.117
3	36	7.70	0.366
4	18	7.52	0.973
5	5	7.56	0.945

The presented p-values reflect Student t-test analyses comparing sucrose concentrations between plants carrying different gene copy numbers. Specifically, we statistically contrasted samples with 1-5 copies against the baseline group (0 copies) to assess dosage-dependent effects on sugar accumulation.

Supplementary Table 14. Summary of the sequencing data for *S. officinarum* XZ and *S. spontaneum* 82-114.

Sample	Sequence Type	Raw data (Gb)	Tissue	coverage
***S. officinarum* XZ**	HiFi	265.46	Leaf	38×
	ONT	84.60	Leaf	12×
	Pore-C	533.64	Leaf	76×
***S. spontaneum* 82-114**	HiFi	230.84	Leaf	32×
	ONT	97.90	Leaf	13×
	Pore-C	297.88	Leaf	41×

Supplementary Table 15. Statistics of contig-level assemblies

Sample	*S. officinarum* XZ	*S. spontaneum* 82-114
Assembly size (Mb)	879.72	766.92
No. of contigs	43	84
Maximum length (bp)	199,350,810	112,080,839
N90 (bp)	70,263,914	55,009,629
N80 (bp)	73,778,006	71,346,441
N70 (bp)	79,020,188	72,848,913
N60 (bp)	81,220,248	74,883,938
N50 (bp)	84,795,178	94,642,250
Average length (bp)	20,458,623.65	9,129,945.57

Supplementary Table 16. BUSCO analysis of genome assembly completeness

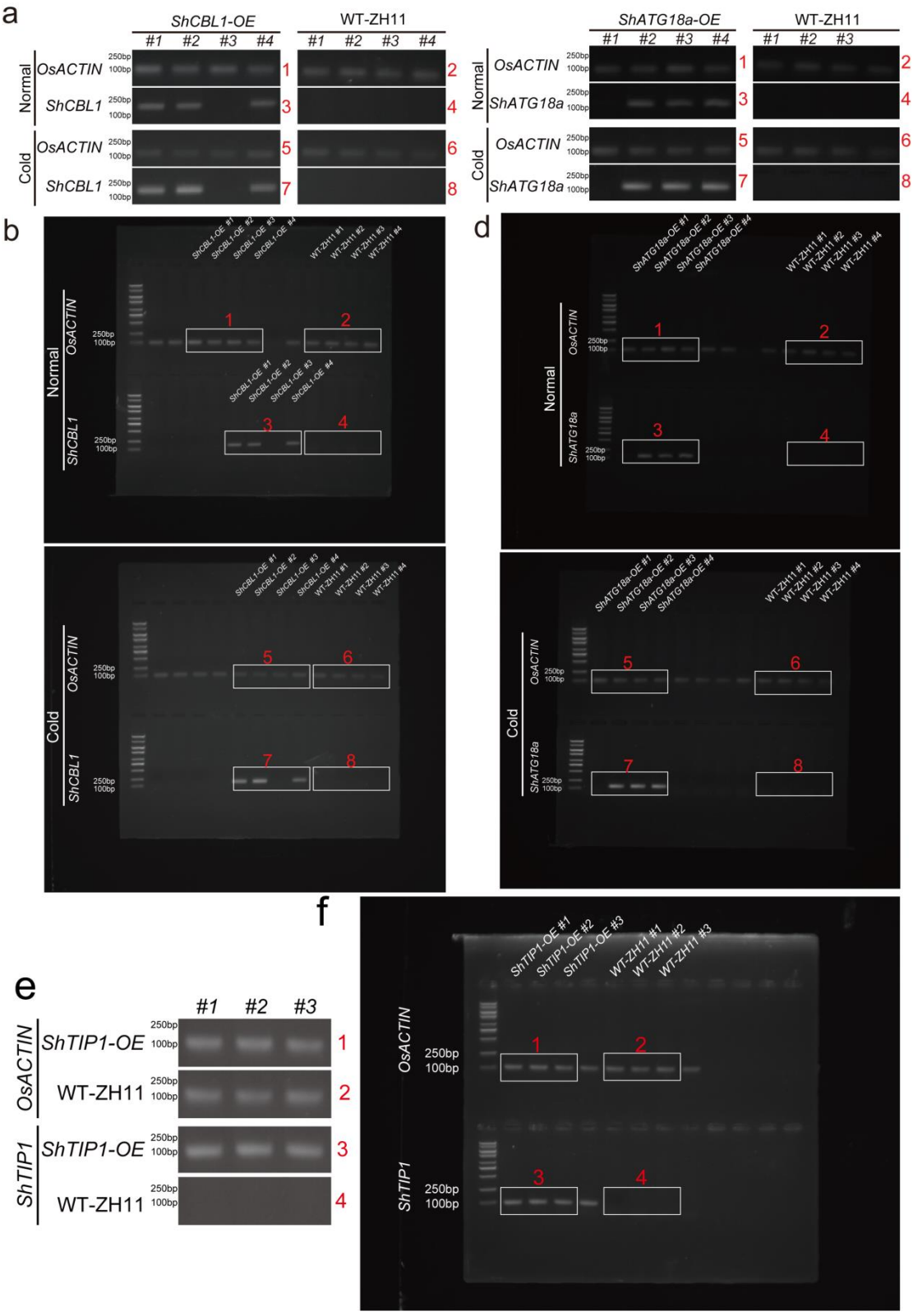
Description	*S. officinarum* XZ		*S. spontaneum* 82-114	
	Number	Percentage (%)	Number	Percentage (%)
Complete BUSCOs (C)	1561	96.7	1543	95.6
Complete and single-copy BUSCOs (S)	1503	93.1	1437	89.0
Complete and duplicated BUSCOs (D)	58	3.6	106	6.6
Fragmented BUSCOs (F)	11	0.7	30	1.9
Missing BUSCOs (M)	42	2.6	41	2.5
Total BUSCO groups searched	1614	100	1614	100

Supplementary Table 17. Overview of chromosome assemblies

Sample	*S. officinarum* XZ	*S. spontaneum* 82-114
Chr01 (bp)	120,506,795	112,080,839
Chr02 (bp)	106,044,810	95,000,000
Chr03 (bp)	103,574,169	84,740,017
Chr04 (bp)	93301000	73,157,053
Chr05 (bp)	85,859,784	102,740,651
Chr06 (bp)	79,020,188	94,642,250
Chr07 (bp)	70263914	83,869,679
Chr08 (bp)	63,437,812	66,201,441
Chr09 (bp)	73,778,006	-
Chr10 (bp)	81,220,248	-
Total Number of contigs	45	89
Total Length of contigs (bp)	879,720,817	766,915,428
Total Length of chromosome level assembly (bp)	877,011,926	712,432,630
Anchor rate (%)	99.69	92.9

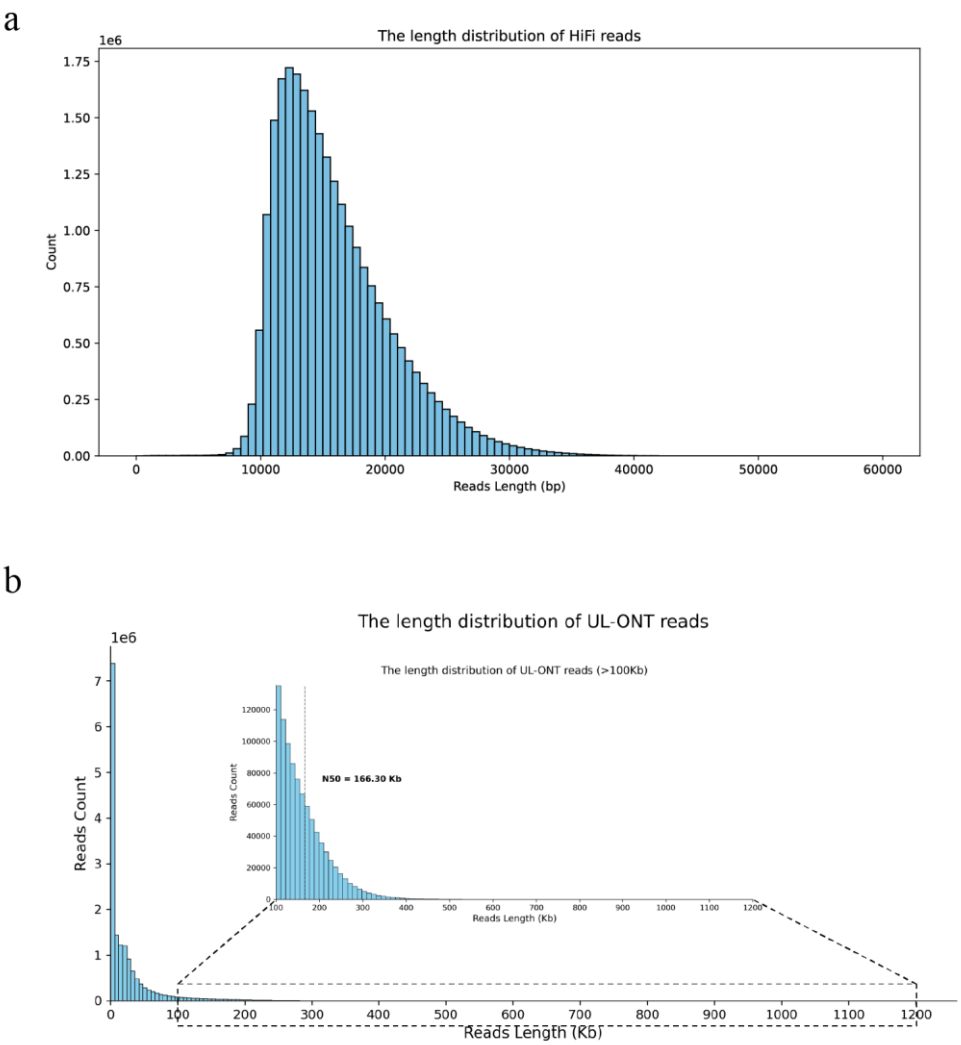
Supplementary Table 18. Statistics for telomeres of the *S. officinarum* XZ and *S. spontaneum* 82-114 genomes. Telomere repeat monomer: AAACCCT.

Sample	Chrid	status	leftnum	leftdirection	rightnum	rightdirection
***S. officinarum* XZ**	Chr01	both	3301	+	595	-
	Chr02	both	18583	+	1262	-
	Chr03	both	14110	+	983	-
	Chr04	both	17058	+	1183	-
	Chr05	both	512	+	909	-
	Chr06	both	610	+	885	-
	Chr07	both	1835	+	872	-
	Chr08	both	412	+	416	-
	Chr09	both	389	+	264	-
	Chr10	both	861	+	273	-
***S. spontaneum* 82-114**	Chr01	both	214	+	114	-
	Chr02	both	345	+	1122	-
	Chr03	both	907	+	300	-
	Chr04	both	242	+	681	-
	Chr05	both	399	+	2339	-
	Chr06	both	325	+	416	-
	Chr07	both	1922	+	1908	-
	Chr08	both	181	+	840	-

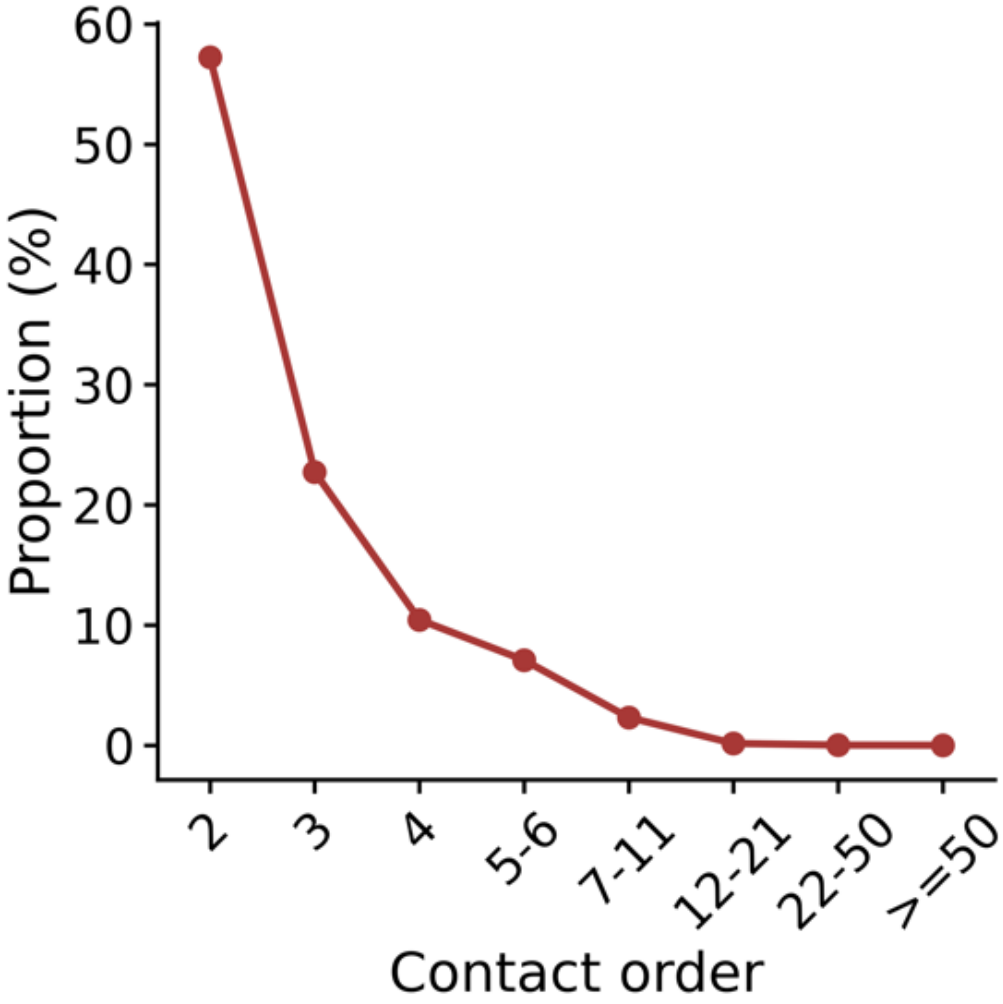


Supplementary Figure 1. Semi-quantitative agarose gel electrophoresis of *ShCBL1*, *ShATG18a*, and *ShTIP1* amplicons in rice (cv. ZH11) overexpression (OE) lines, with corresponding uncropped source gels. (**a,b**) *ShCBL1*: cropped gel (as in Extended Data

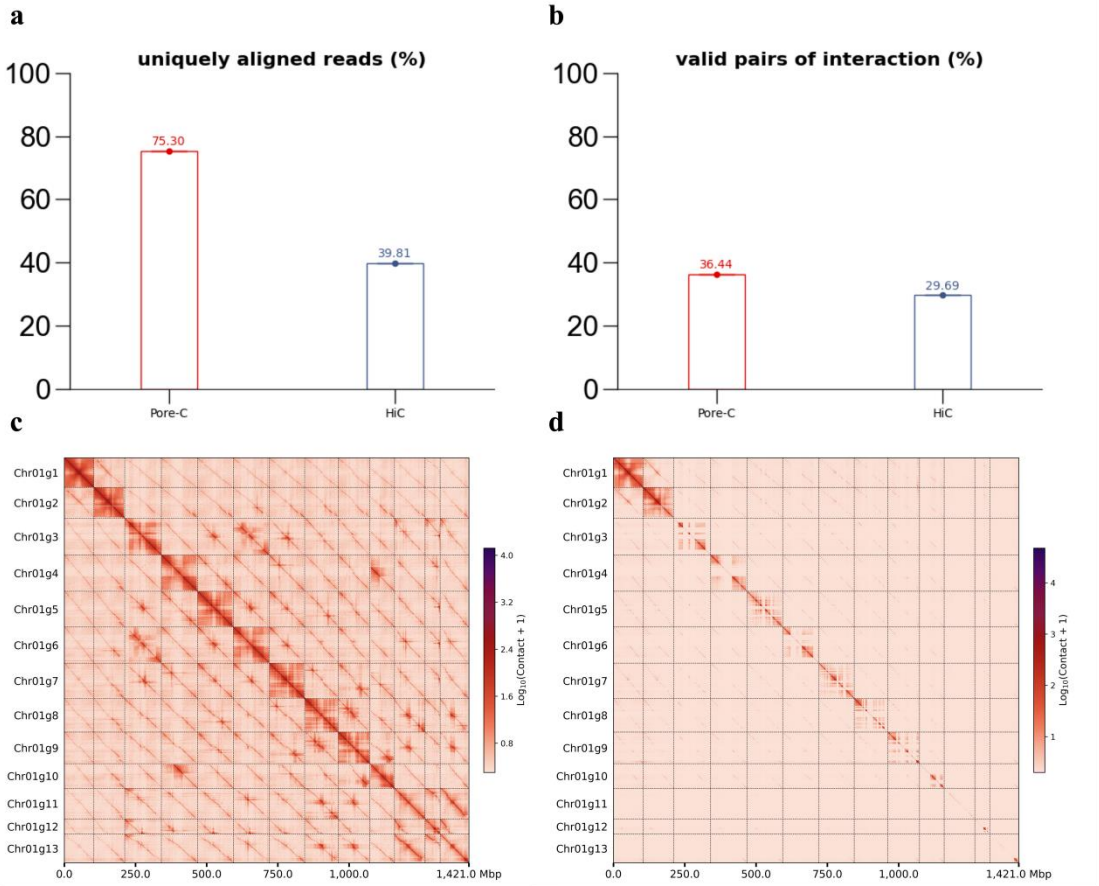
Fig. 5g) and matching uncropped scan comparing OE lines with WT-ZH11. (**c,d**) ShATG18a: cropped gel (as in Extended Data Fig. 5h) and matching uncropped scan comparing OE lines with WT-ZH11. In (b,d), white boxes indicate the regions shown in (a,c); identical red numerals mark the same lanes in cropped and uncropped images. (e,f) *ShTIP1*: cropped/annotated gel (as in Extended Data Fig. 6k) with *OsACTIN* loading control and the corresponding uncropped scan; the white box in (f) marks the cropped region in (e) and red numerals denote matching lanes. *OsACTIN* was used as the loading control, and OsACTIN and target amplicons were run on the same gel.



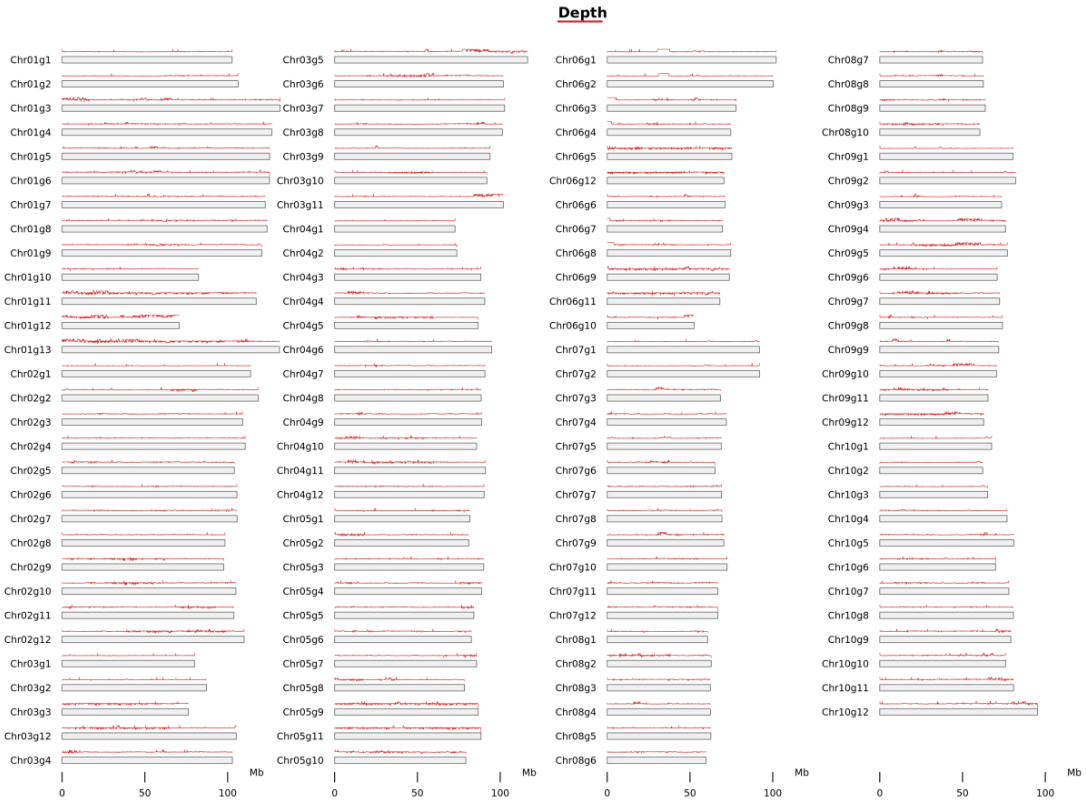
Supplementary Figure 2. Length distribution of sequencing reads. (a) Length distribution of PacBio HiFi reads. (b) Length distribution of ONT ultra-long reads. The distribution of ONT ultra-long reads exceeding 100 kb (used for assembly) is specifically magnified, with the N50 position explicitly indicated.



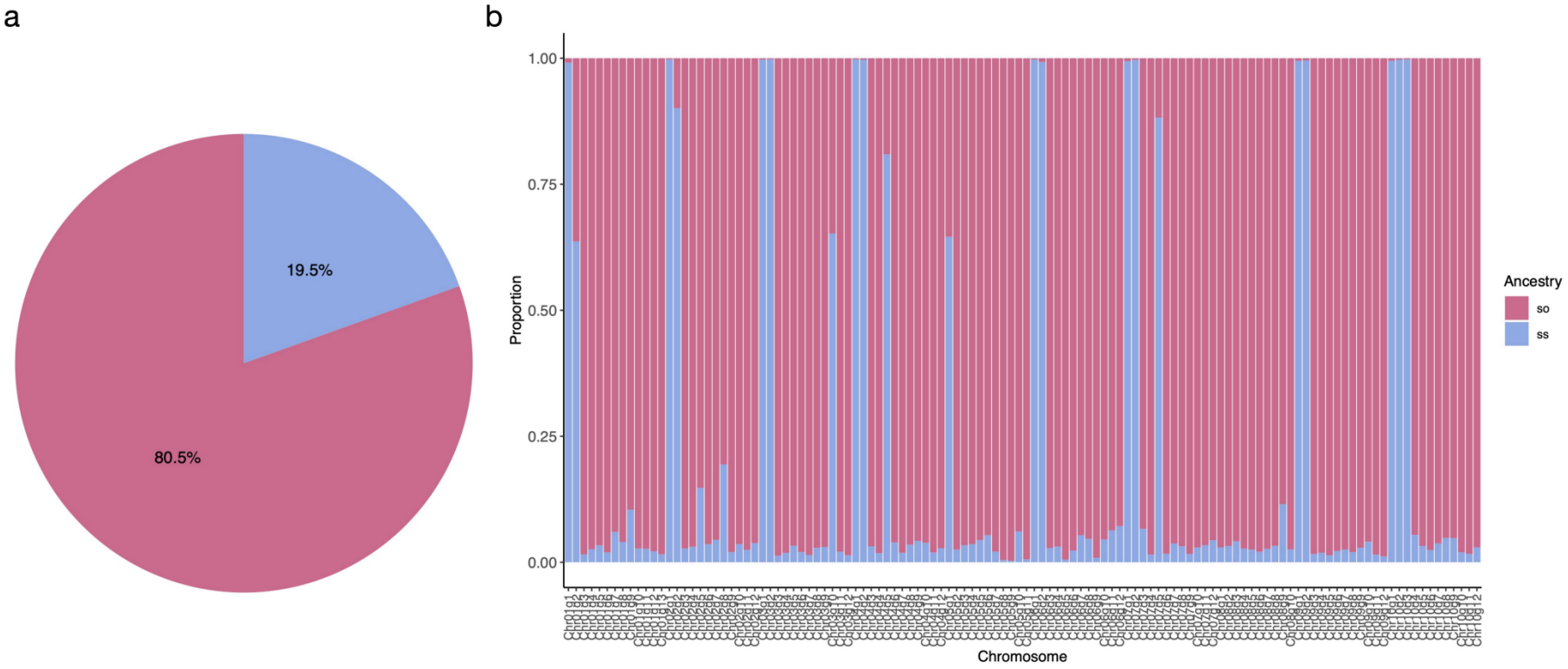
Supplementary Figure 3. Distribution of two-order and high-order chromatin interaction in Pore-C reads. Only Pore-C reads with contact orders higher than two were included in the calculation.



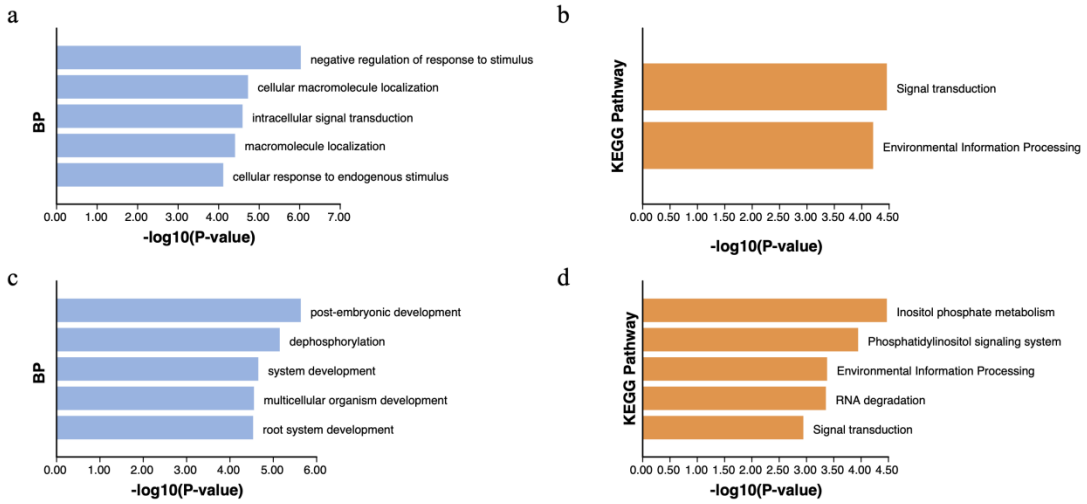
Supplementary Figure 4. Comparison between Pore-C and Hi-C reads (HiC-Pro). (a) the ratio of uniquely mapped reads; (b) the ratio of valid pairs of interaction. chromatin contact heatmaps based on Pore-C data (c) and Hi-C data (d).



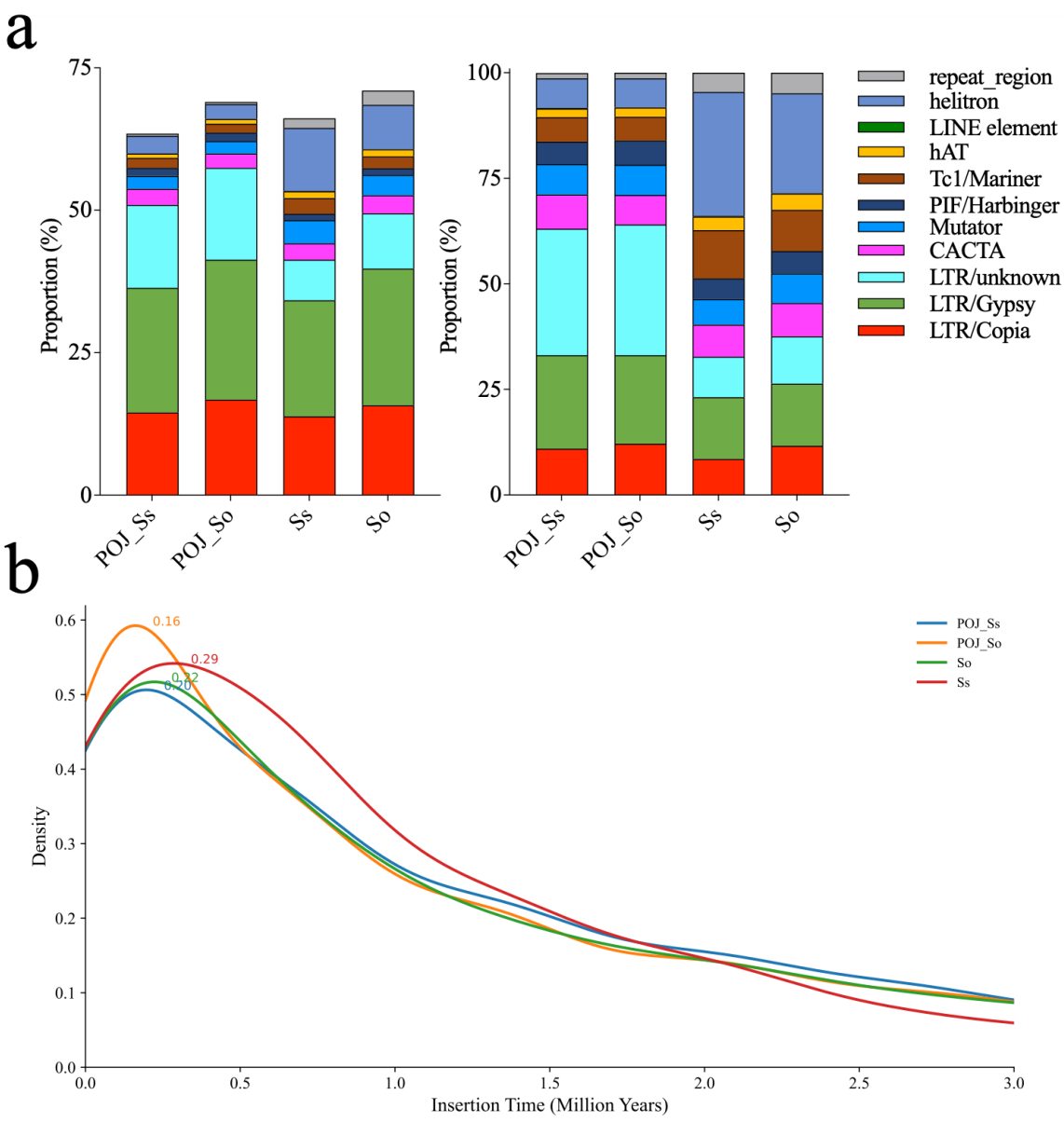
Supplementary Figure 5. Distribution of read depth across the 118 chromosomes based on short-read sequencing.



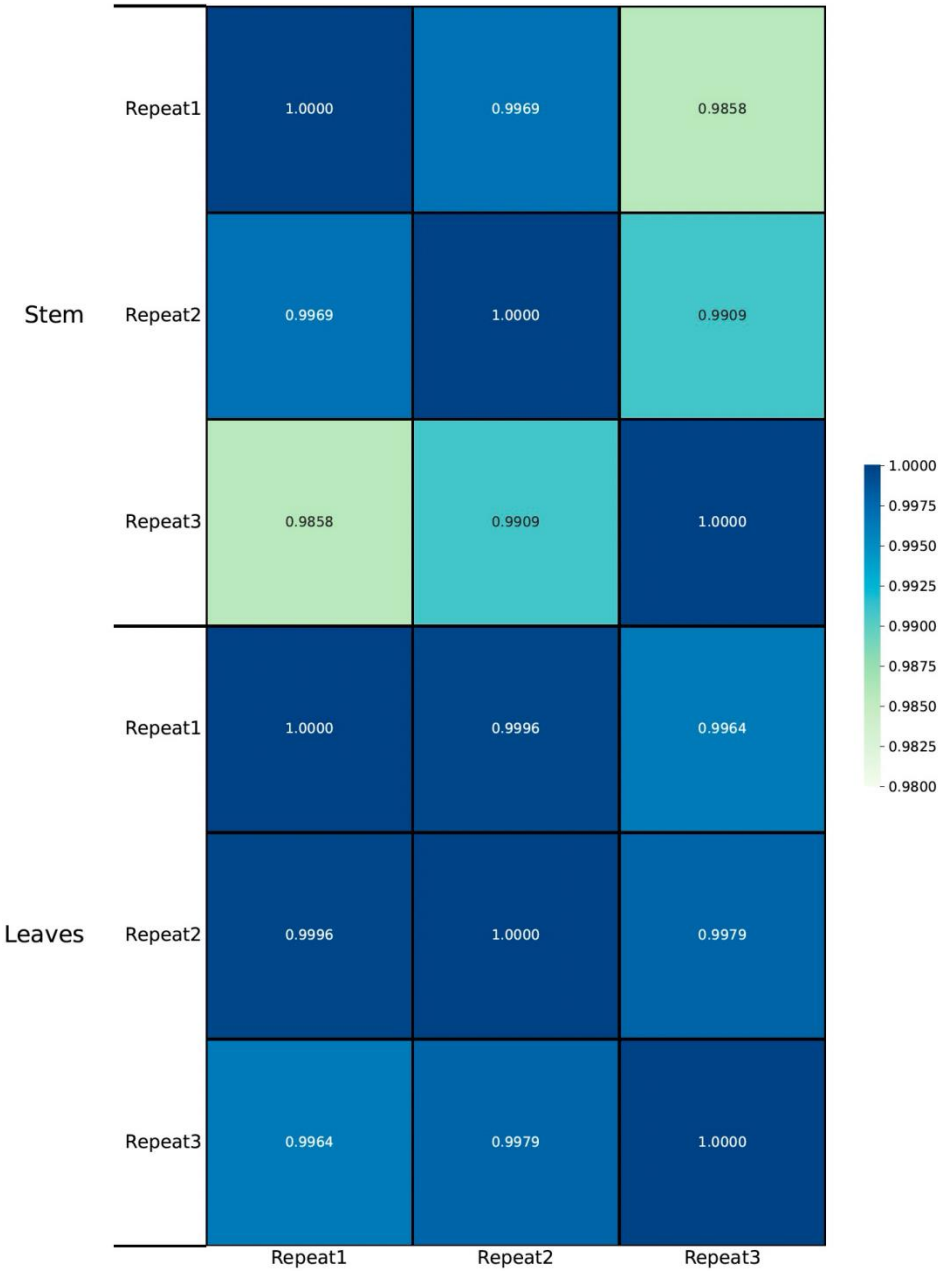
Supplementary Figure 6. Proportion of SS- and SO-derived genomic sequences in the whole-genome (a) and 118 chromosomes (b).



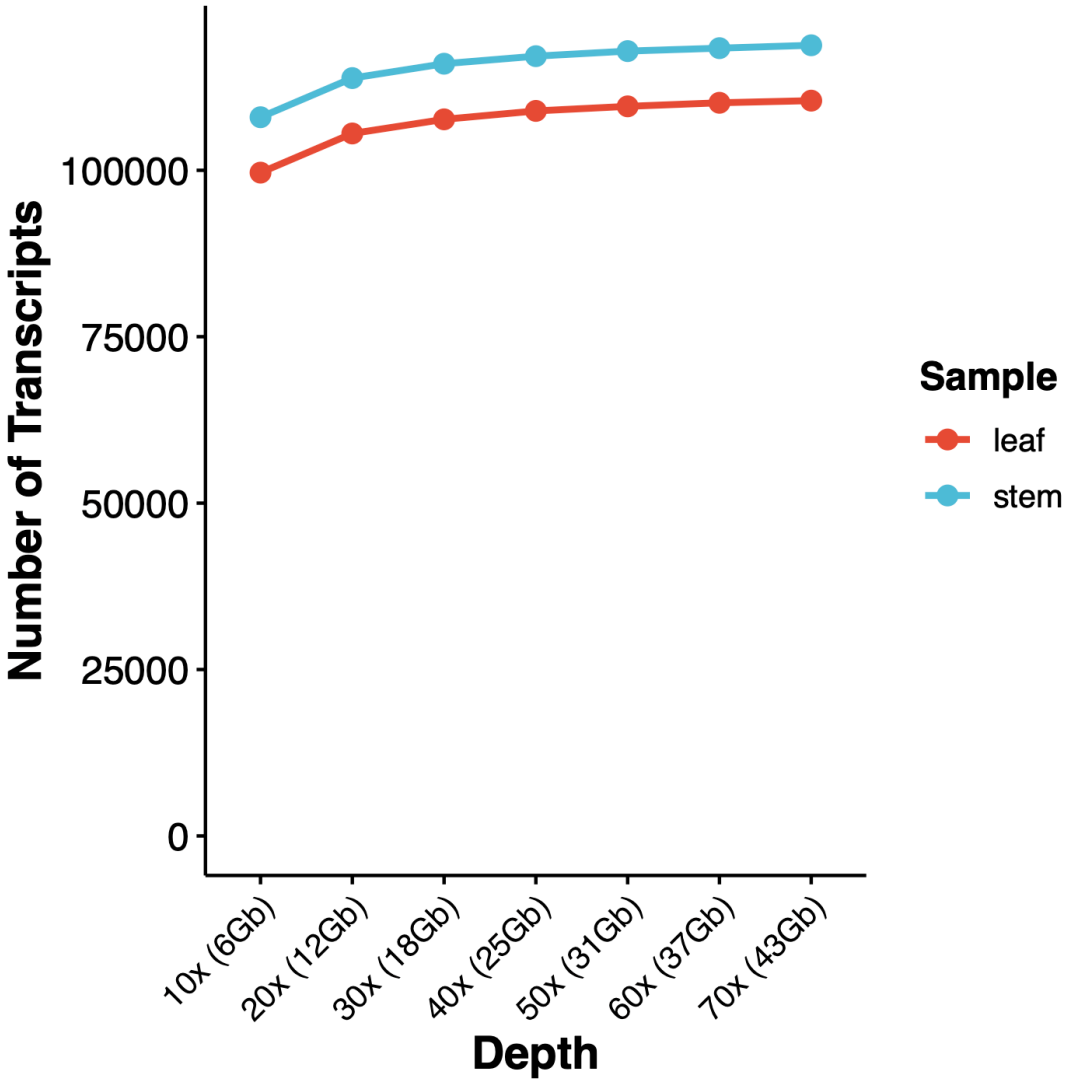
Supplementary Figure 7. Functional annotation of homoeologous genes between two subgenomes. (a) GO Enrichment Analysis of Biological Processes in HE on SS Chromosomes; (b) KEGG Enrichment Analysis of HE on SS Chromosomes; (c) GO Enrichment Analysis of Biological Processes in HE on SO Chromosomes. (d) KEGG Enrichment Analysis of HE on SO Chromosome



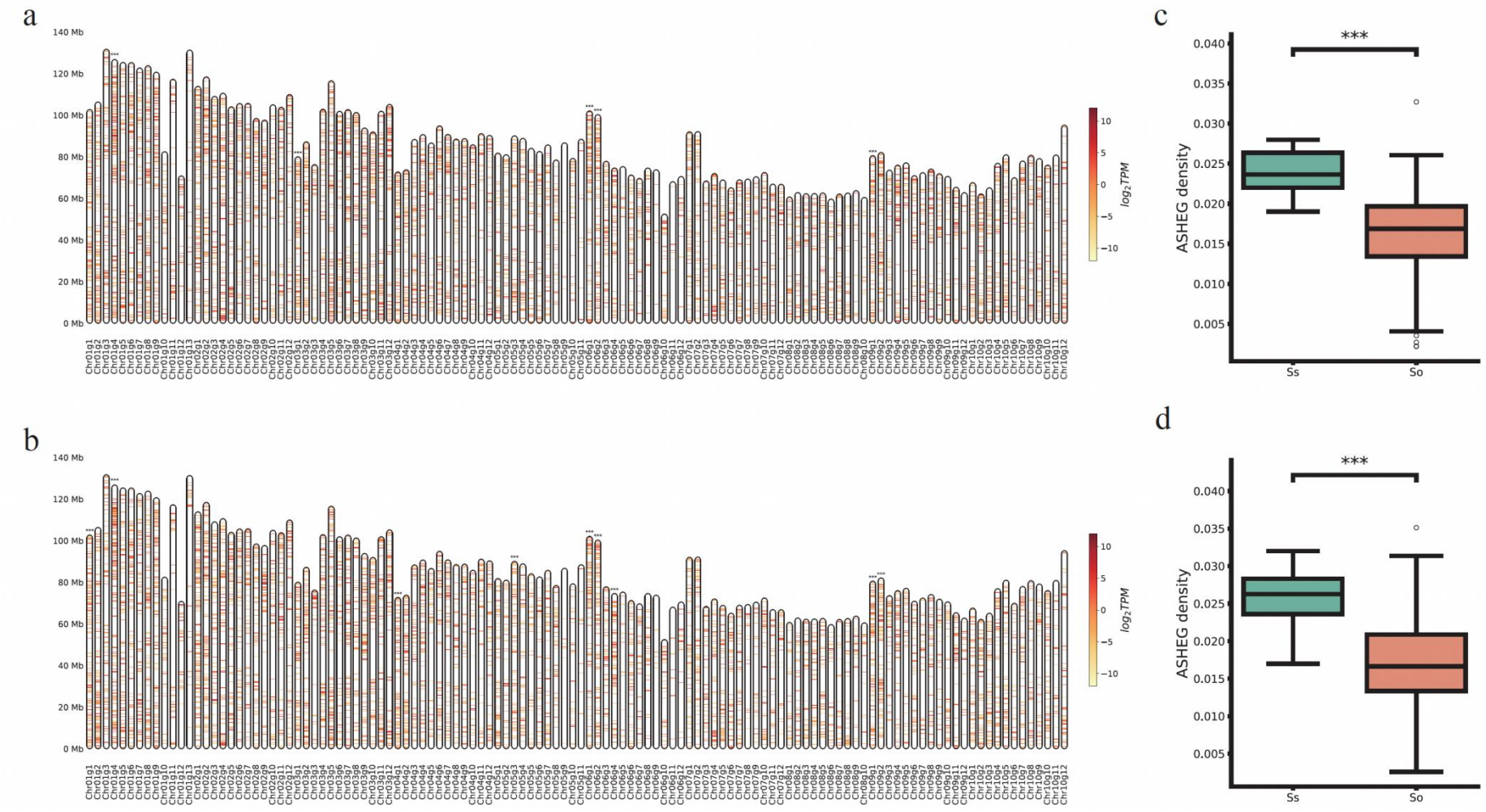
Supplementary Figure 8. TE composition and insertion times of LTR retrotransposons of POJ2878 subgenomes compared with two ancestral genomes. a. Stacked histograms representing the contribution of each TE superfamily to the two POJ2878 subgenomes and two ancestral species genomes, the left shows the composition of different types of TEs. The right shows the proportion of different types of TEs; **b.** Analysis of intact LTR insertion times in the two of POJ2878 subgenomes and two ancestral special genomes. The horizontal and vertical axes represent the insertion time of intact LTR retrotransposons and the density of intact LTR retrotransposons, respectively. The value above this peak indicates the estimated insertion time for each set of LTRs.



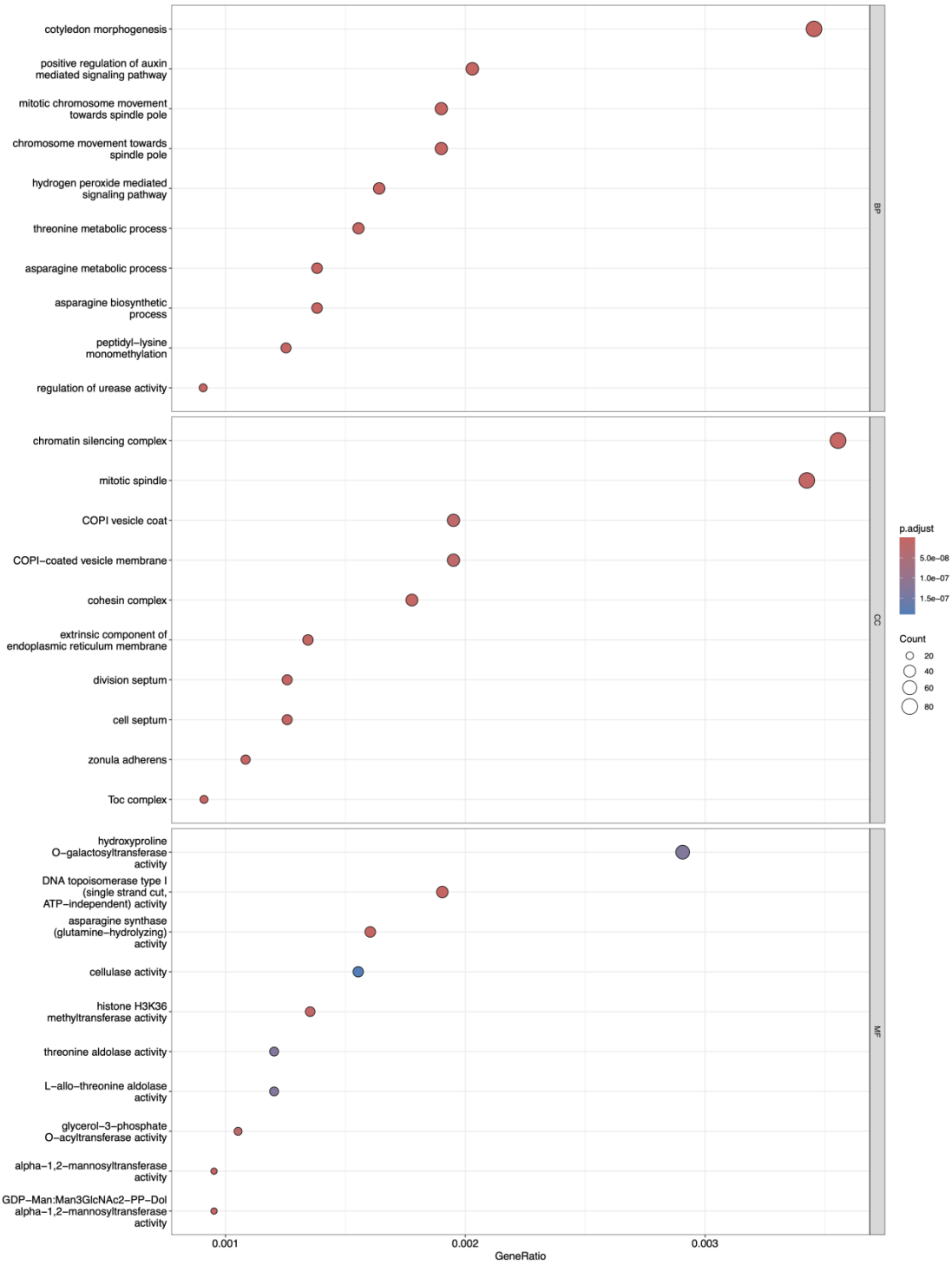
Supplementary Figure 9. Pearson's correlation coefficient of gene expression across replicates using MAS-ISO-seq data



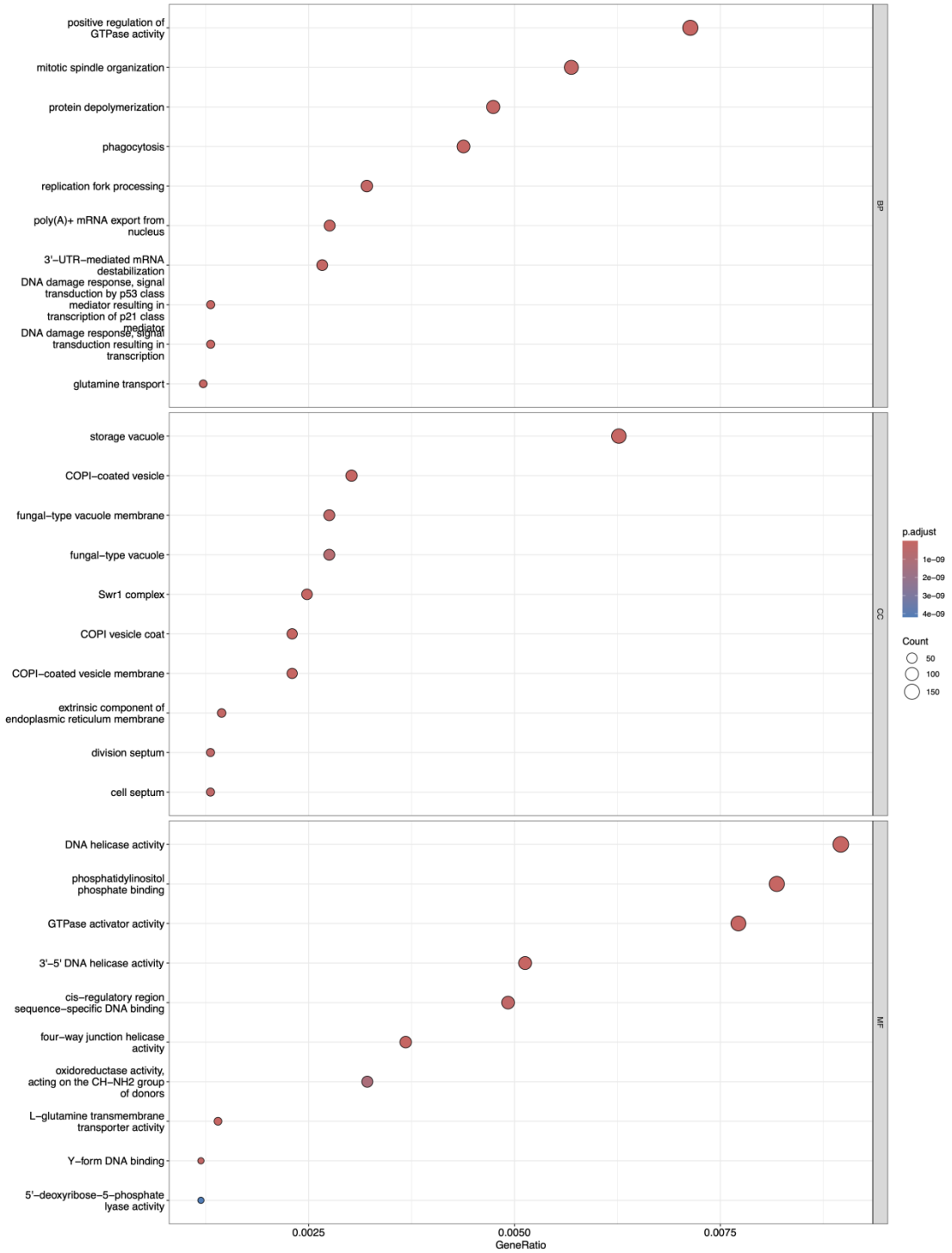
Supplementary Figure 10. Saturation analysis of MAS-ISO-seq transcript discovery. The saturation curves illustrate the number of unique transcripts identified (y-axis) as a function of sequencing depth (x-axis) in leaf (red) and stem (blue) samples. The plateauing of both curves indicates that the sequencing depth was adequate for comprehensive transcriptome coverage, capturing the majority of expressed transcripts in both tissues.


a
b
c
d
140 Mb
120 Mb
100 Mb
80 Mb
60 Mb
40 Mb
20 Mb
0 Mb
log2TPM
ASHEG density
Ss
So
***

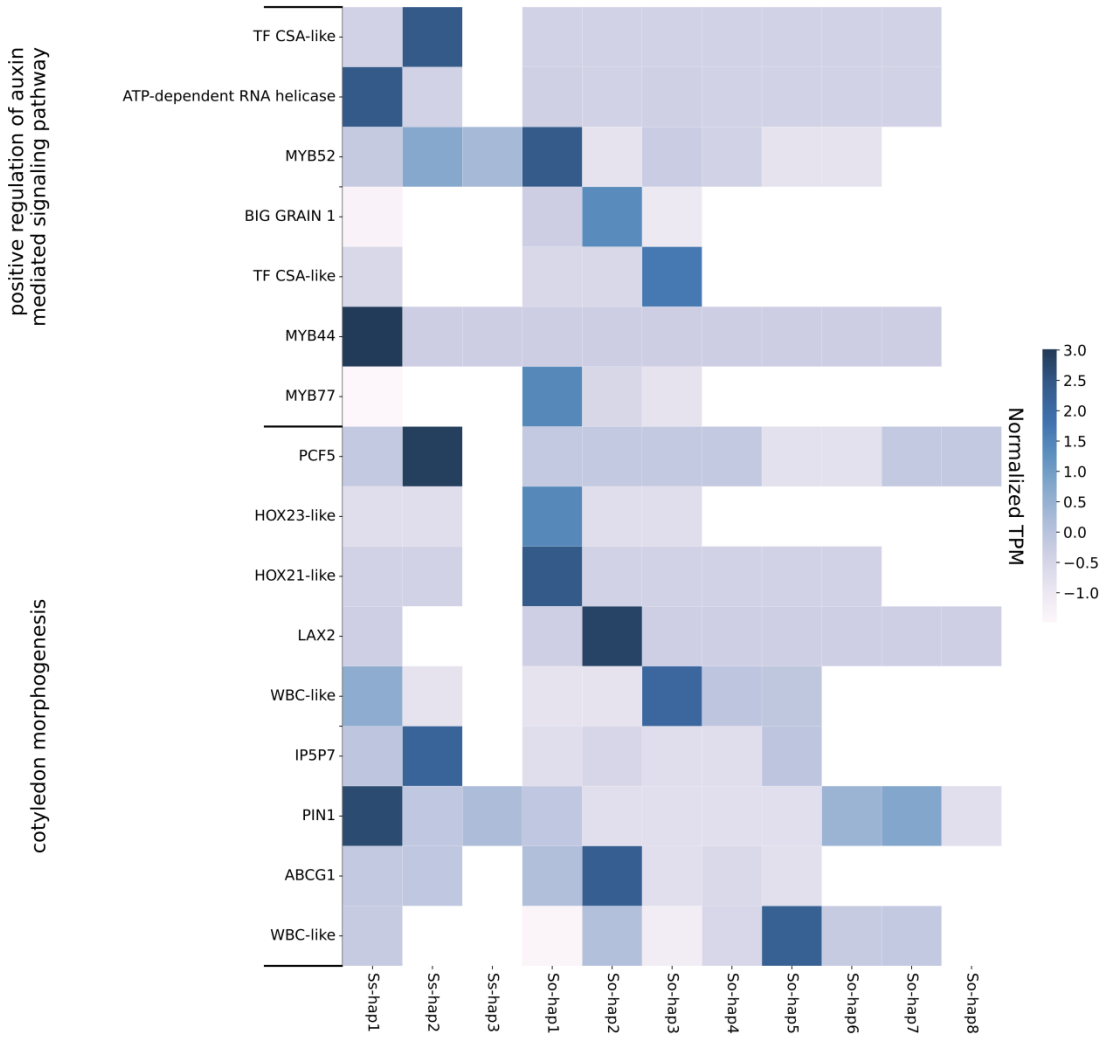
Supplementary Figure 11. Chromosome-level distribution and subgenome-biased expression of ASHEGs. (a-b) Distribution of ASHEGs across all 118 chromosomes of POJ2878 in stem tissue (a) and leaf tissue (b). Each vertical bar represents an individual chromosome, with the height indicating chromosome length (Mb). The colored segments within each bar represent the distribution of ASHEGs along the chromosome, with color intensity reflecting the $\log_2$TPM expression level. Chromosome labels on the x-axis indicate chromosome identity. ASHEG density (the proportion of ASHEGs relative to total genes per chromosome) was calculated for each chromosome, and chromosomes with significant ASHEG enrichment were identified using a binomial test followed by Bonferroni correction for multiple testing (adjusted $P < 0.01$). c, d Comparison of ASHEG density between Ss and So subgenomes in stem tissue (c) and leaf tissue (d). Box plots display median (center line), interquartile range (box), 1.5× interquartile range (whiskers), and outliers (circles). Statistical analysis was performed using Mann-Whitney U test (exact $P = 2.45 \times 10^{-7}$ for c, and $P = 2.03 \times 10^{-7}$ for d).



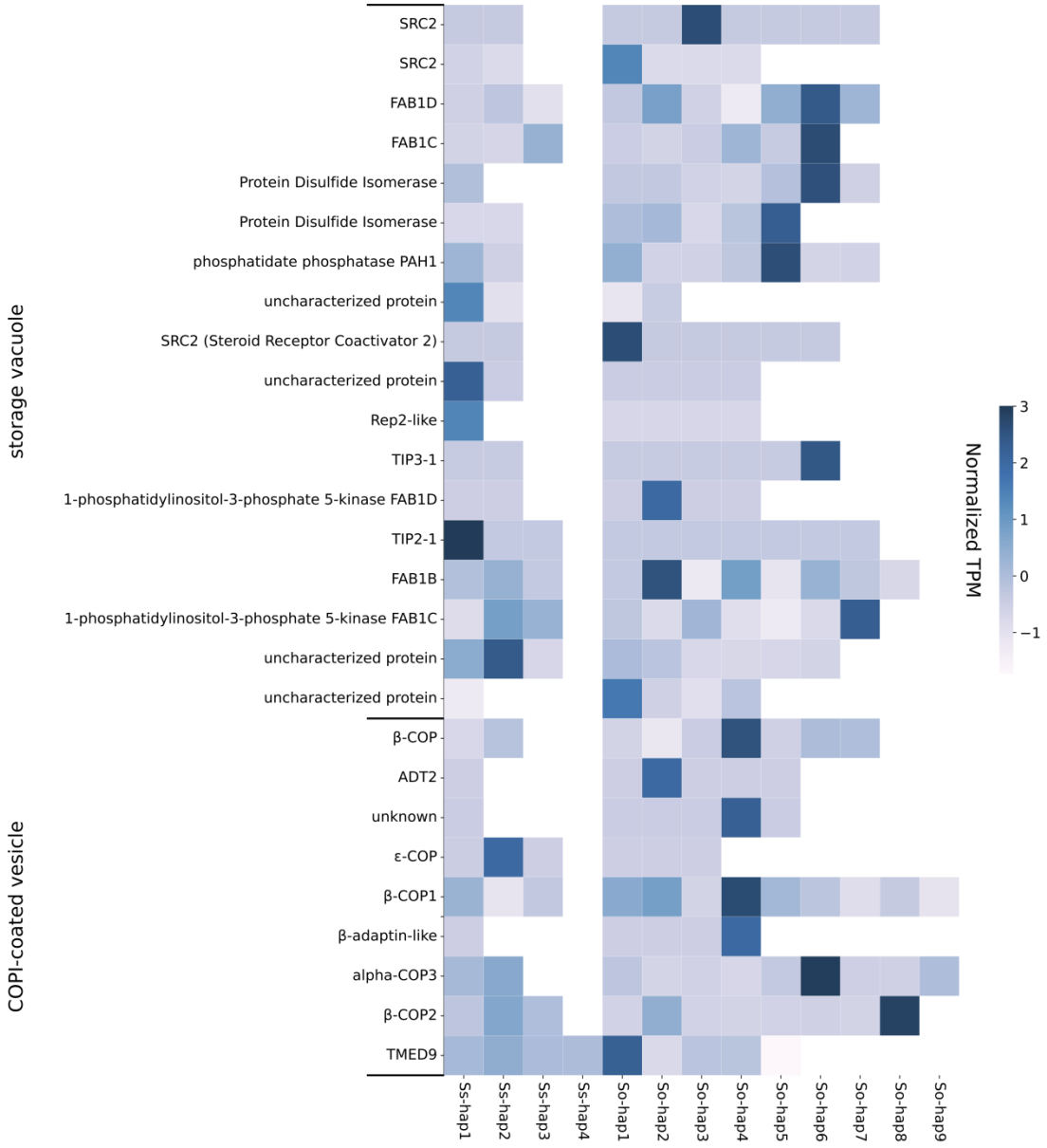
Supplementary Figure 12. GO enrichment analysis of allele-specific highly expressed genes in the mature leaf tissue of POJ2878.



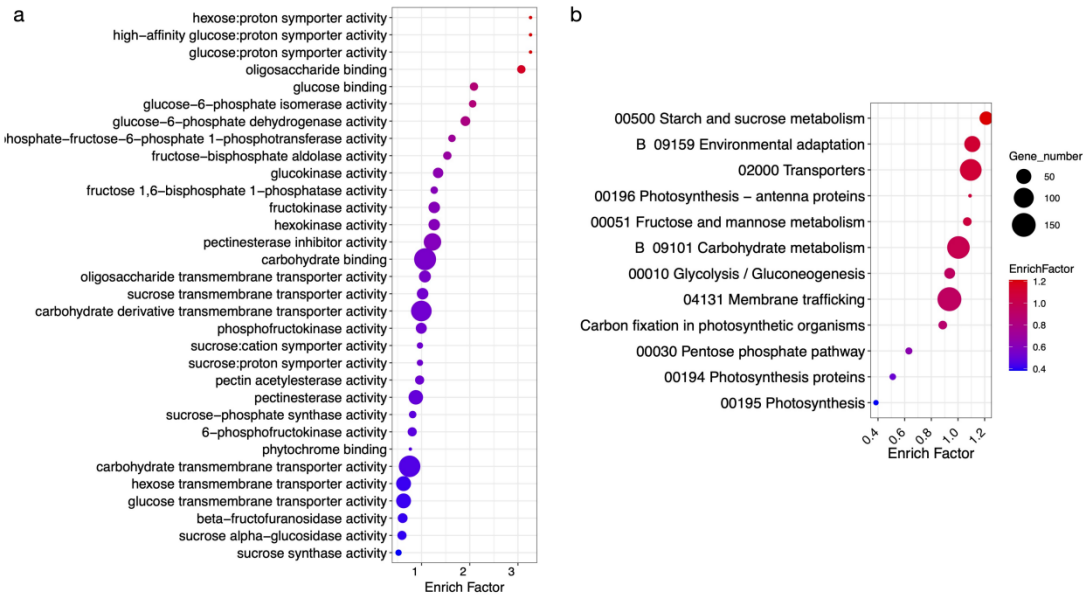
Supplementary Figure 13. GO enrichment analysis of haplotype-specific highly expressed genes in the mature stem tissue of POJ2878.



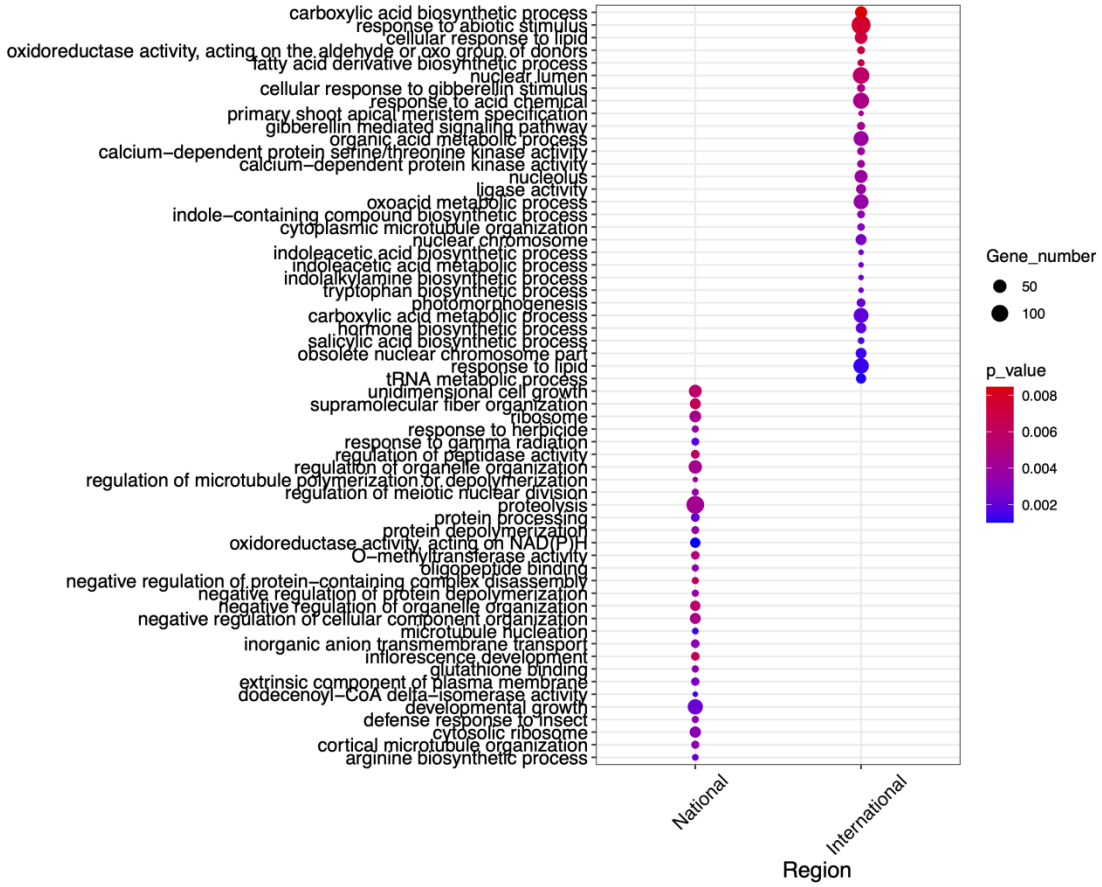
Supplementary Figure 14. The expression of haplotype-specific highly expressed genes that are significantly enriched in cotyledon morphogenesis and positive regulation of auxin-mediated signaling pathways in POJ2878 leaf tissue. The TPM values were normalized using z-score.



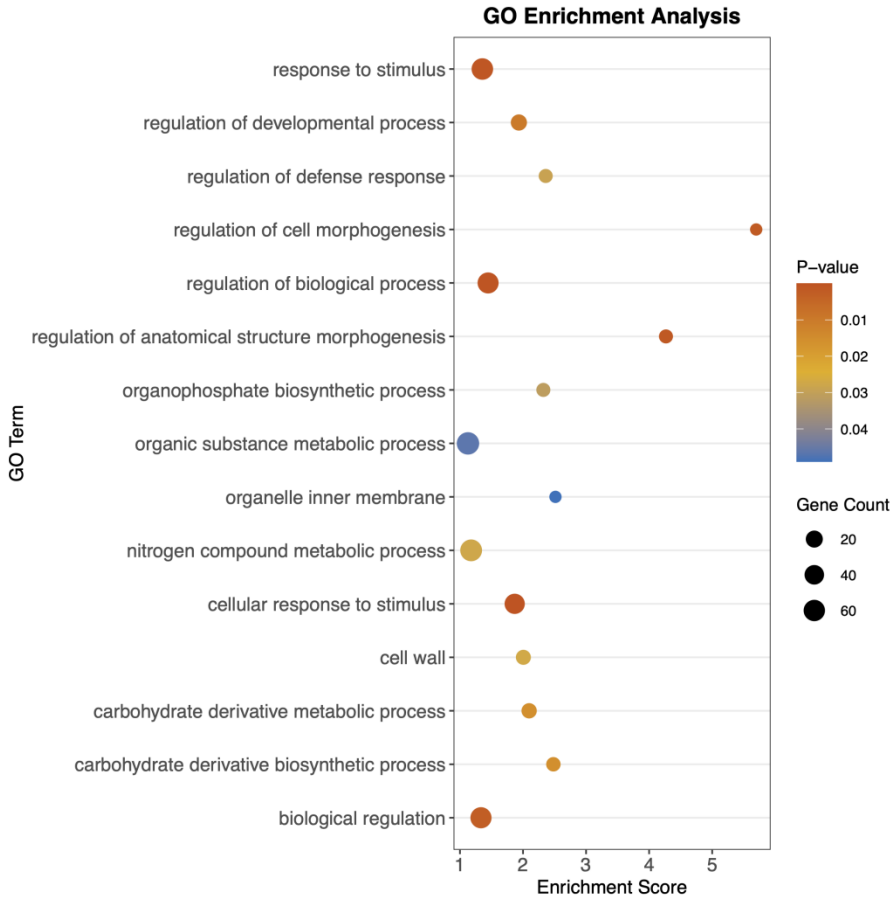
Supplementary Figure 15. The expression of haplotype-specific highly expressed genes that are significantly enriched in storage vacuole and COPI-coated vesicle terms in POJ2878 mature stem tissue. The TPM values were normalized using z-score.



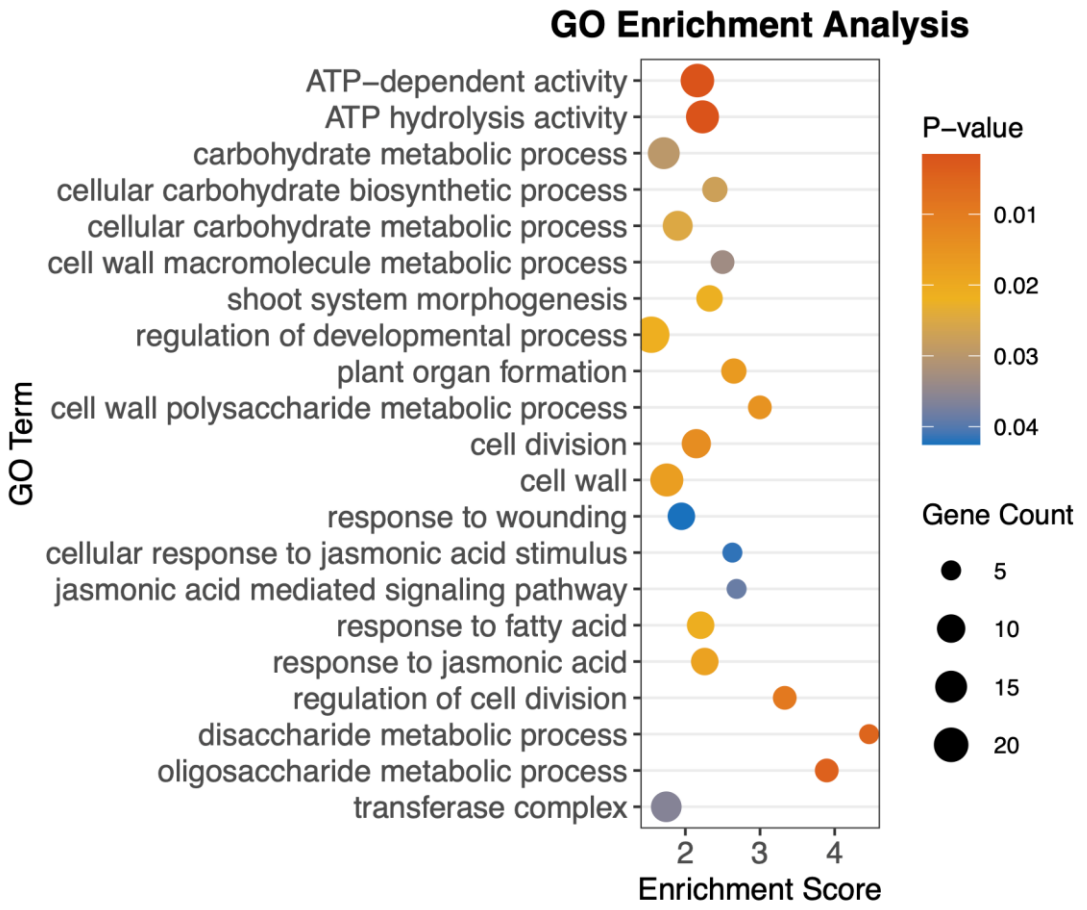
Supplementary Figure 16. GO (a) and KEGG (b) enrichment analysis of genes within favorable haplotypes that have a high-level of IBD density.



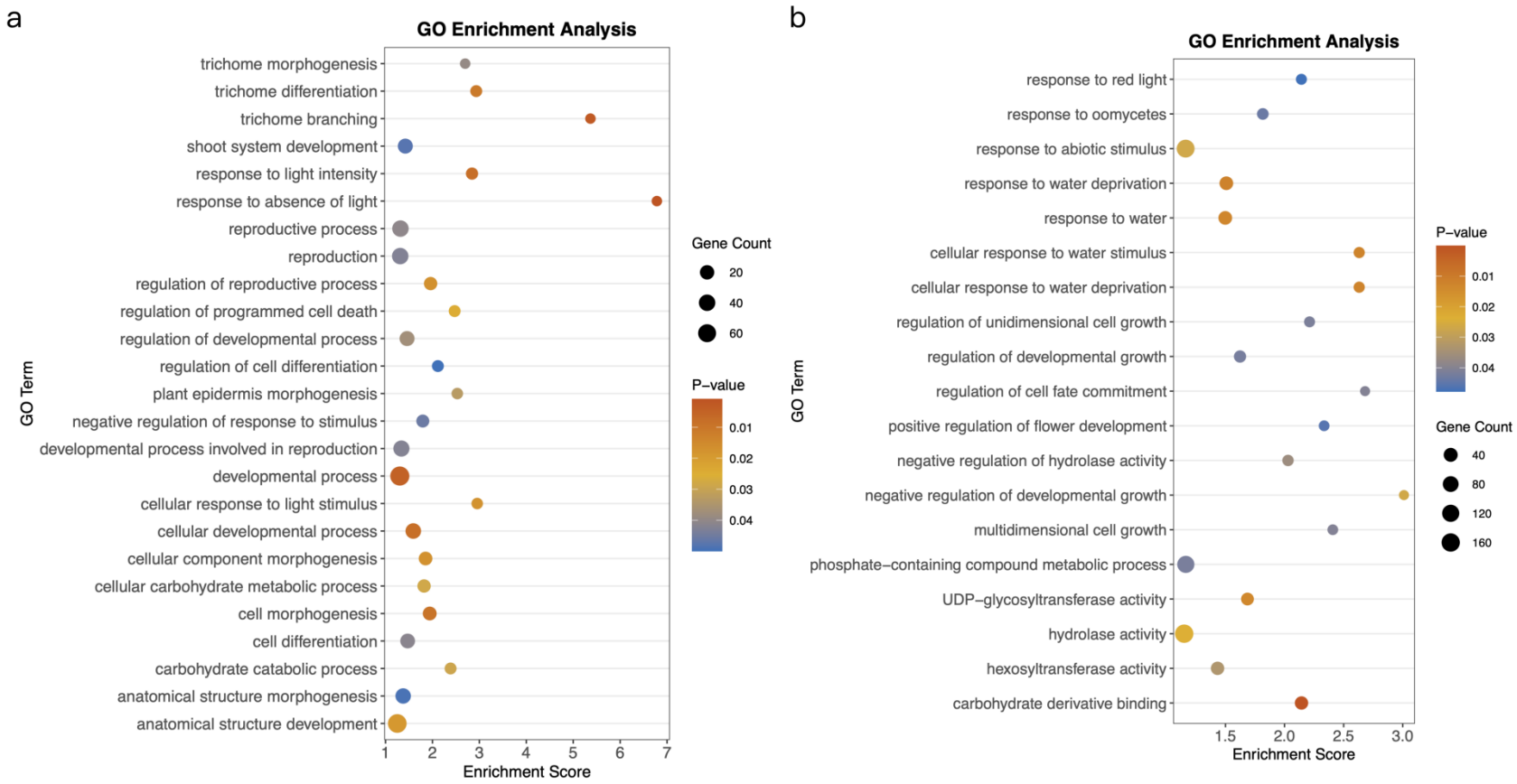
Supplementary Figure 17. GO and KEGG enrichment analysis of favorable haplotypes in Chinese and non-Chinese cultivars.



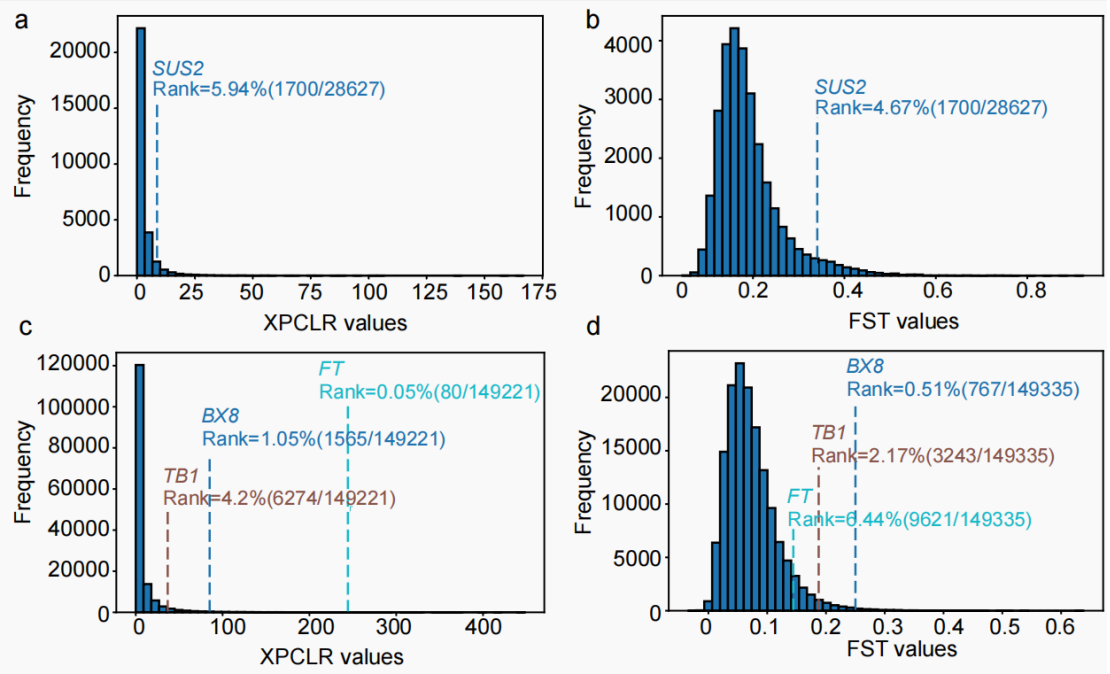
Supplementary Figure 18. GO enrichment of the selective swept genes under natural selection in *S. spontaneum* population.



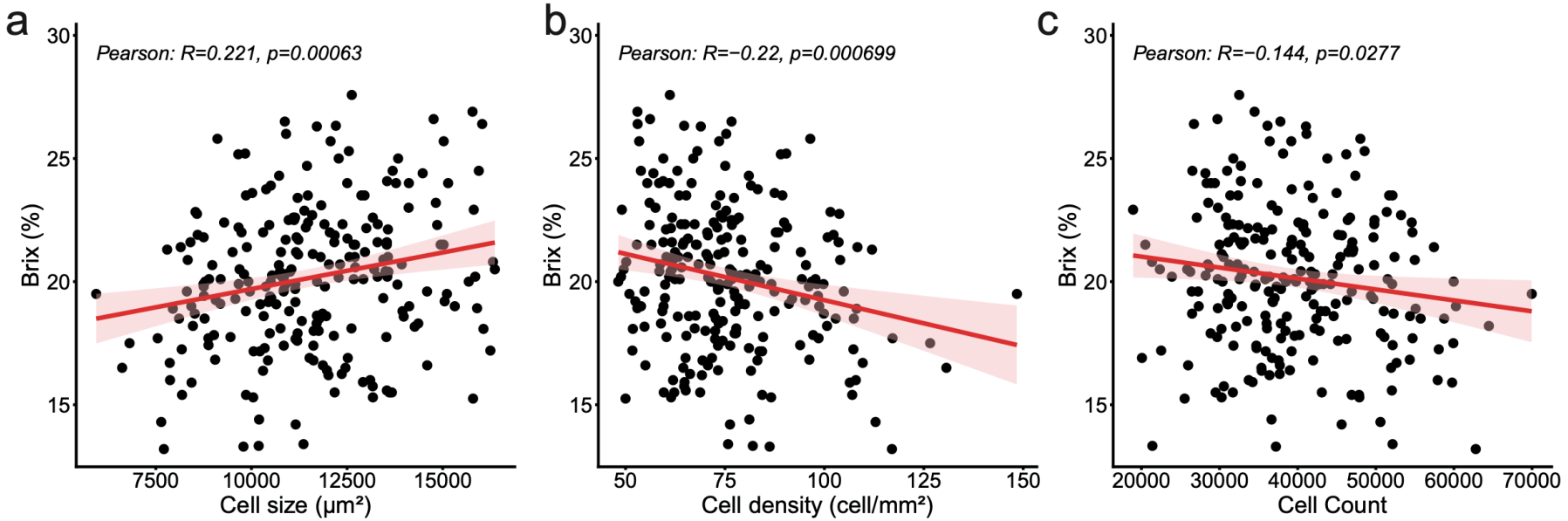
Supplementary Figure 19. GO enrichment analysis of genes under selective sweeps during domestication in the *S. officinarum* population compared to the *S. robustum* population.



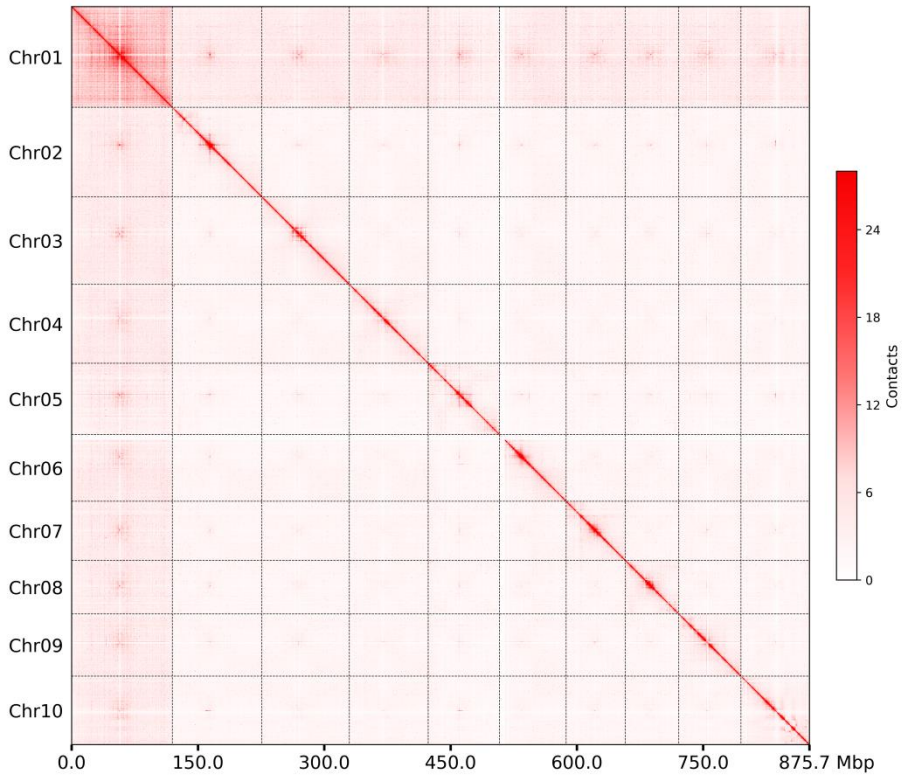
Supplementary Figure 20. GO enrichment analysis of genes under selective sweeps during domestication in the *S. hybrid* population. This analysis involves in comparison of *S. hybrid* population with *S. spontaneum* population (a) and with *S. officinarum* population (b).



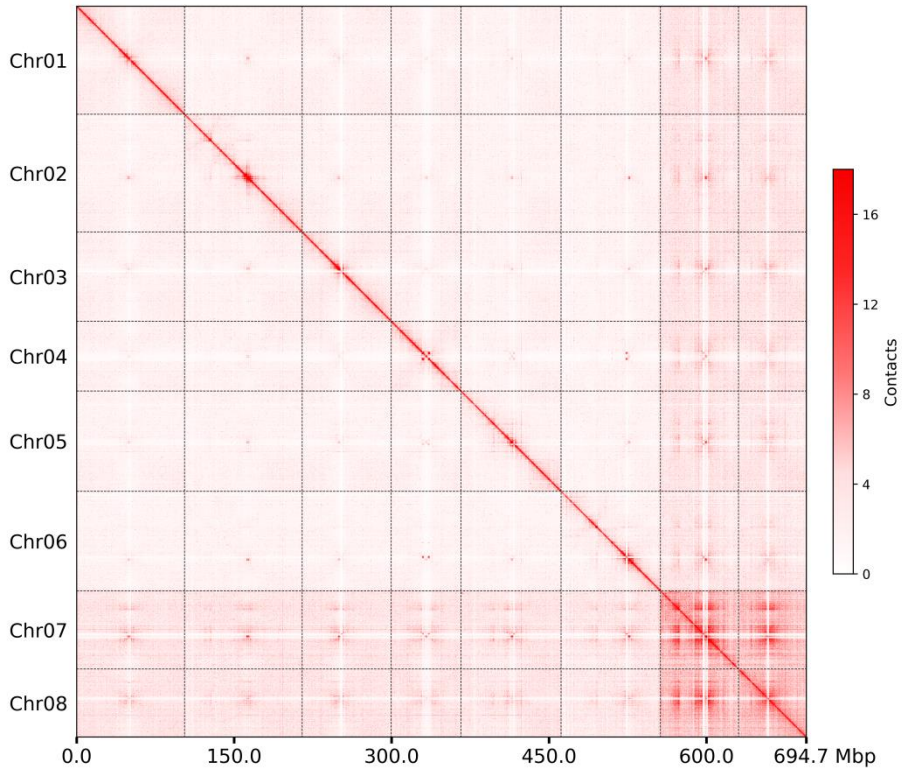
Supplementary Figure 21: Empirical distributions of XP-CLR and F_{ST} values for each gene region. (a-b) Rankings of *SUS2* (Cul vs Ss). (c-d) Rankings of TB1, BX8, and FT (Cul vs So).



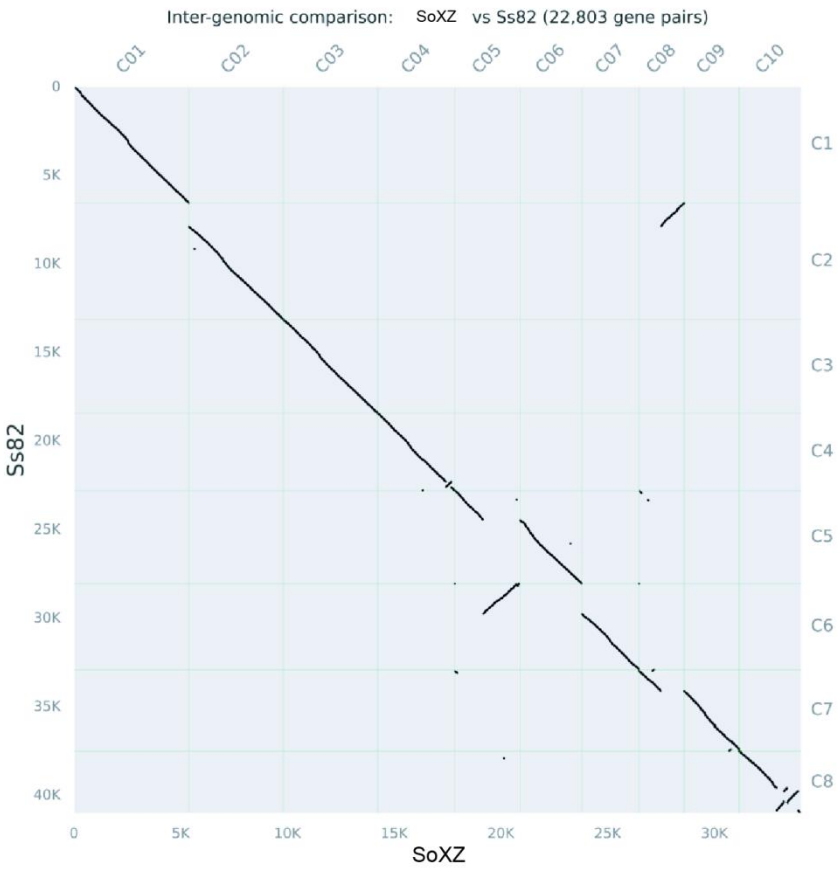
Supplementary Figure 22. The correlation analysis between Brix and parenchyma cell traits. Scatter plots showing the relationships between Brix content (%) and three parenchyma cell characteristics: (a) cell size (μm²), (b) cell density (cell/mm²), and (c) cell count. Each black dot represents an individual sample. The total sample size (n) is 265. Red lines indicate linear regression trends with shaded areas representing 95% confidence intervals. Pearson correlation coefficients (R) and p-values are displayed in each panel.



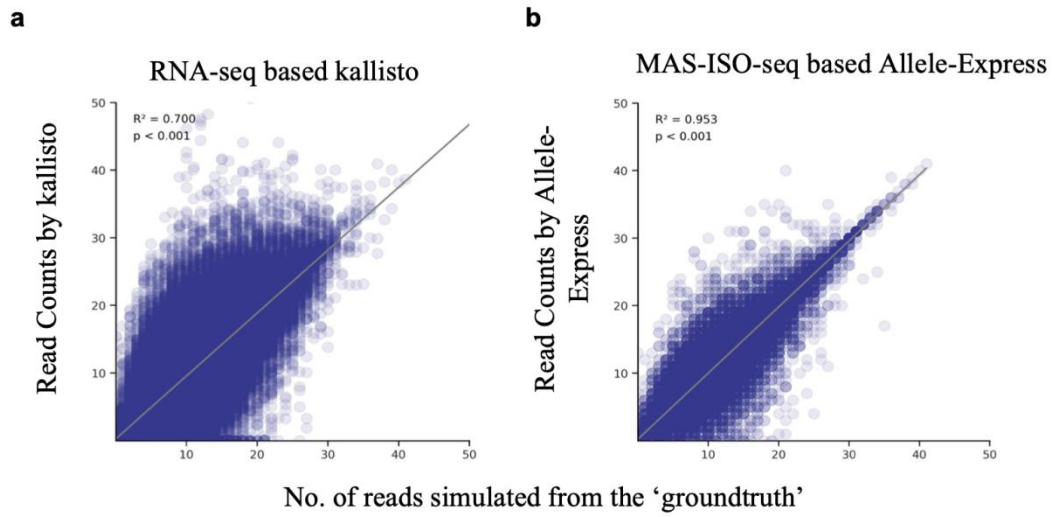
Supplementary Figure 23. Pore-C interaction heatmaps of *S. officinarum* XZ



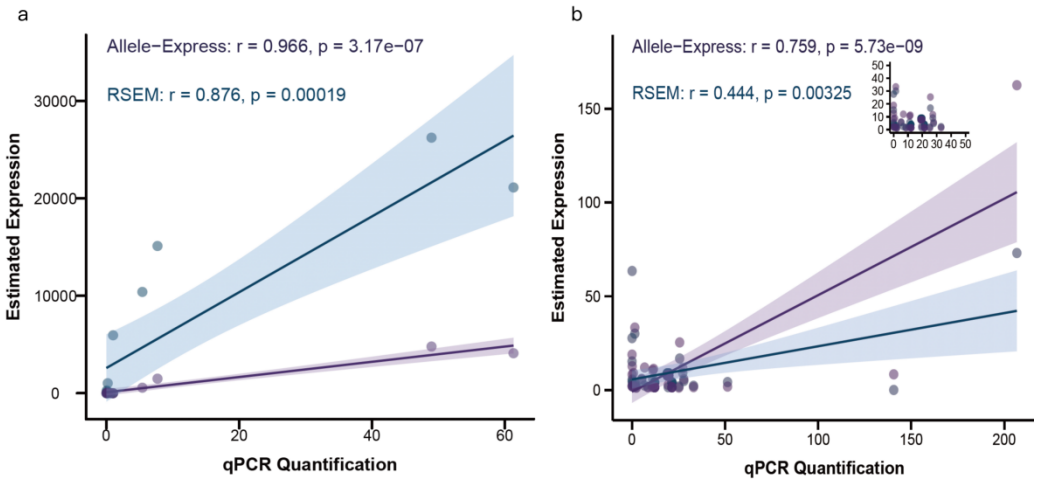
Supplementary Figure 24. Pore-C interaction heatmaps of *S. spontaneum* 82-114 genomes.



Supplementary Figure 25. Dotplot between *S. officinarum* XZ and *S. spontaneum* 82-114.



Supplementary Figure 26. Benchmarking of different gene quantification approaches in the polyploid sugarcane genome. The x-axis represents the number of reads simulated from the 'ground truth' based on the POJ2878 genome. The y-axis shows the calculated read counts from RNA-seq using the Kallisto algorithm (a) and MAS-ISO-Seq using the Allele-Express algorithm (b).



Supplementary Figure 27. Experimental validation of Allele-Express performance using quantitative RT-PCR. Correlation analysis between computational gene expression estimates and experimental qRT-PCR measurements across different genomic complexities. (a) Validation in diploid Arabidopsis thaliana showing correlation between qRT-PCR quantification (x-axis) and estimated expression levels (y-axis) for multiple selected genes. (b) Validation in polyploid sugarcane genome showing the same comparison. Shaded areas represent 95% confidence intervals for the regression lines. The inset in panel (b) shows a zoomed view of the low-expression region. Statistical significance was determined using Pearson correlation analysis.