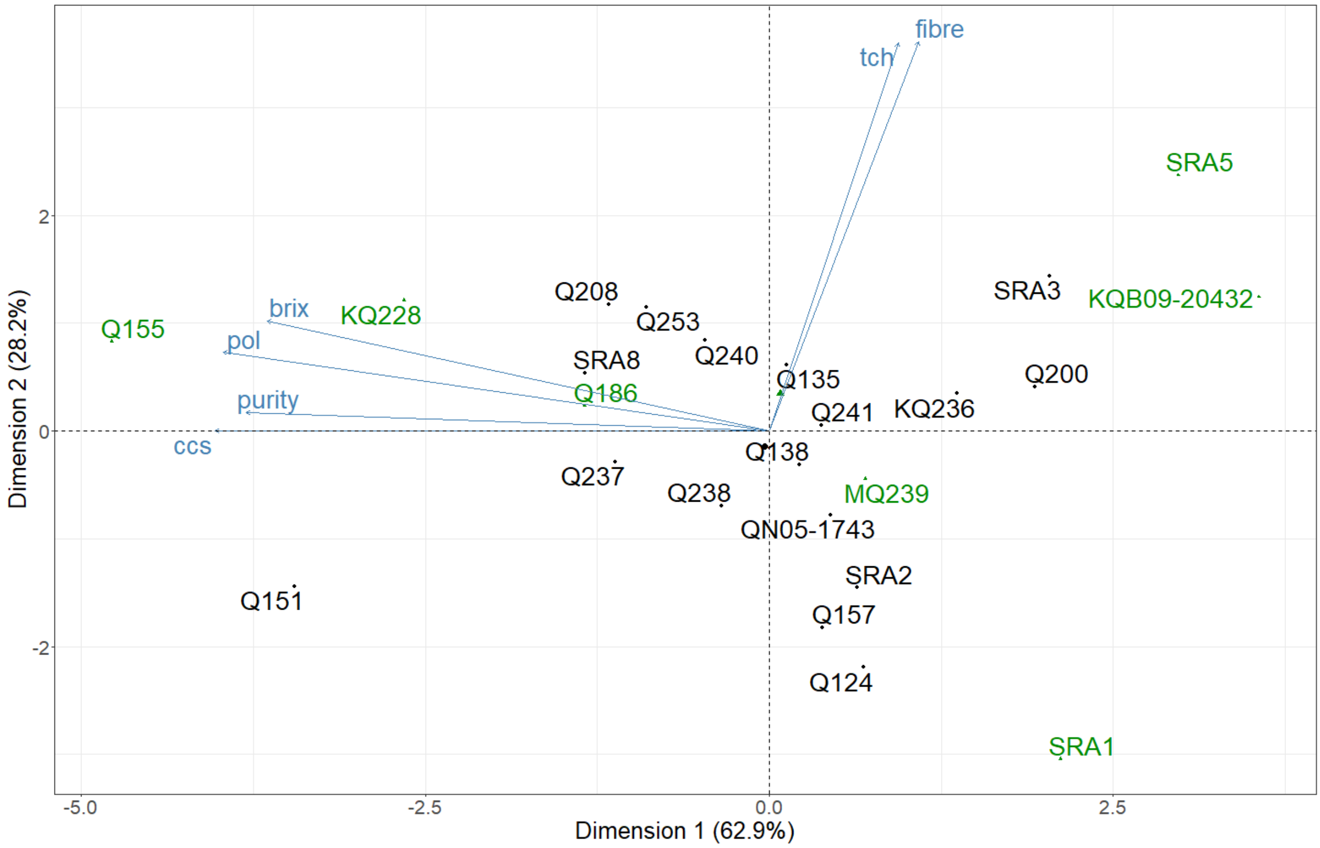
Limited allele-specific gene expression in highly polyploid sugarcane

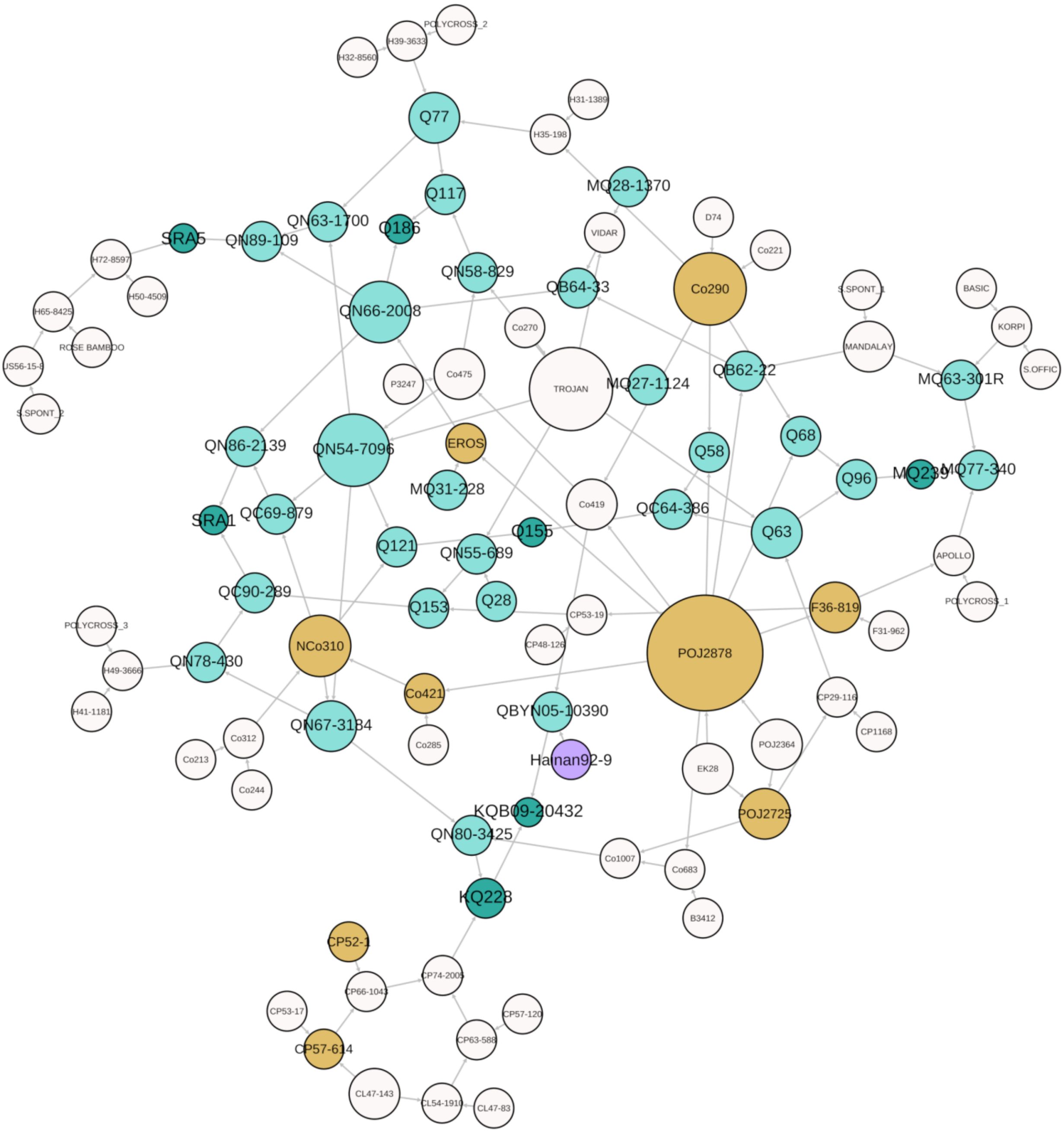
Gabriel Rodrigues Alves Margarido, Fernando Henrique Correr, Agnelo Furtado, Frederik C Botha and Robert James Henry

Supplemental Figures

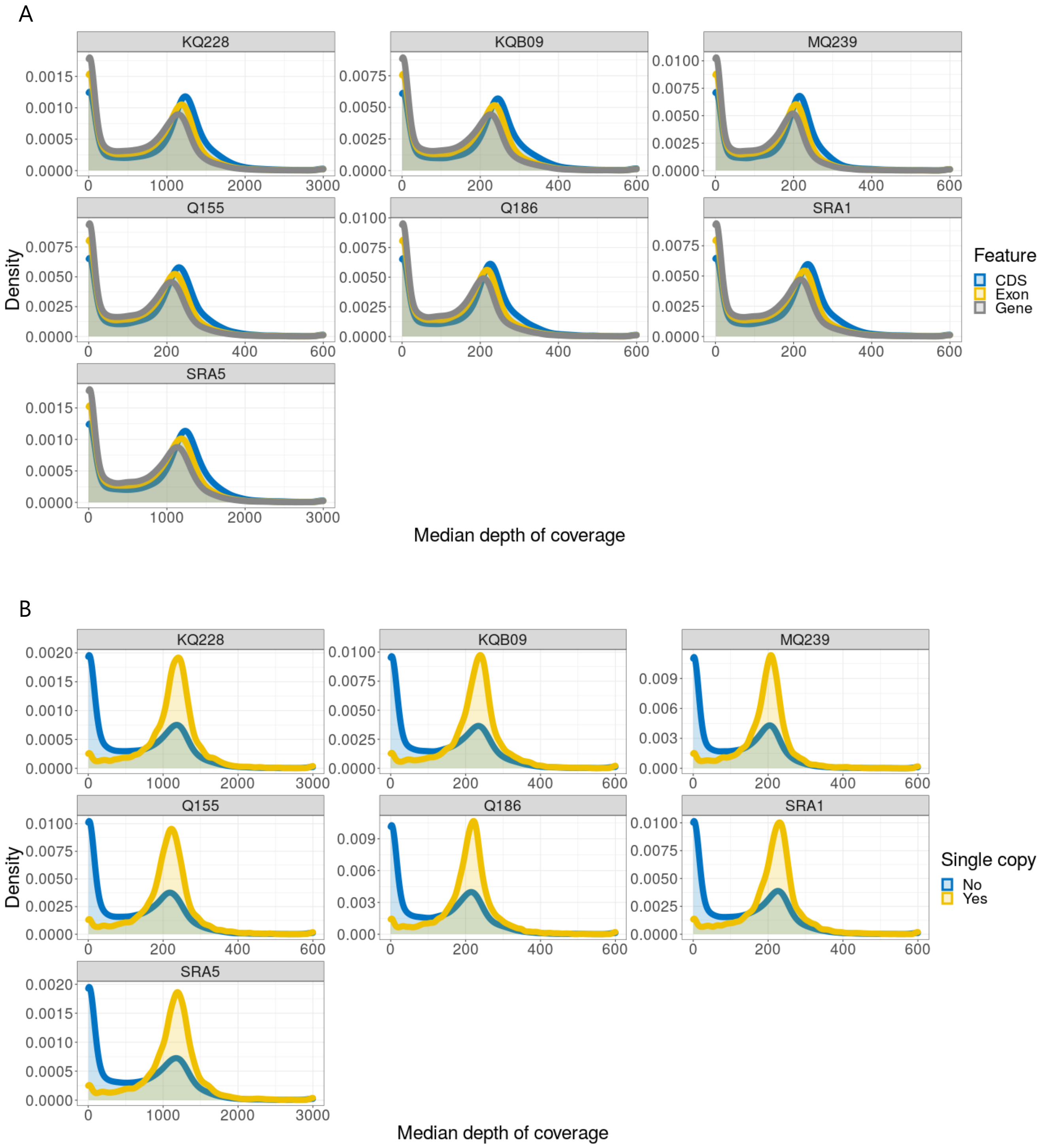
Supplemental tables and methods are available in separate files.



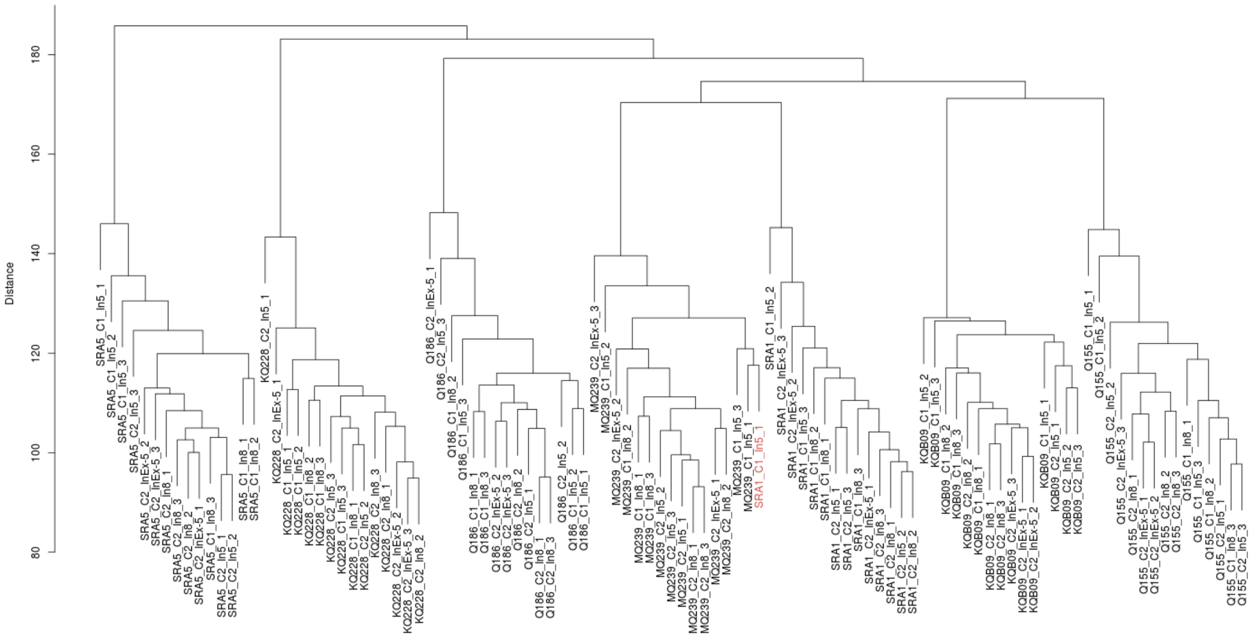
Supplemental Fig S1. Principal Component Analysis of the phenotypic variation observed in sugarcane hybrids. A total of 24 genotypes were evaluated in field conditions for yield (tonnes of cane per hectare, tch), fibre content, soluble solids (brix), sucrose content (polarity, pol), sucrose purity and commercial cane sugar (ccs). Genotypes marked in green were used to study allele-specific expression.



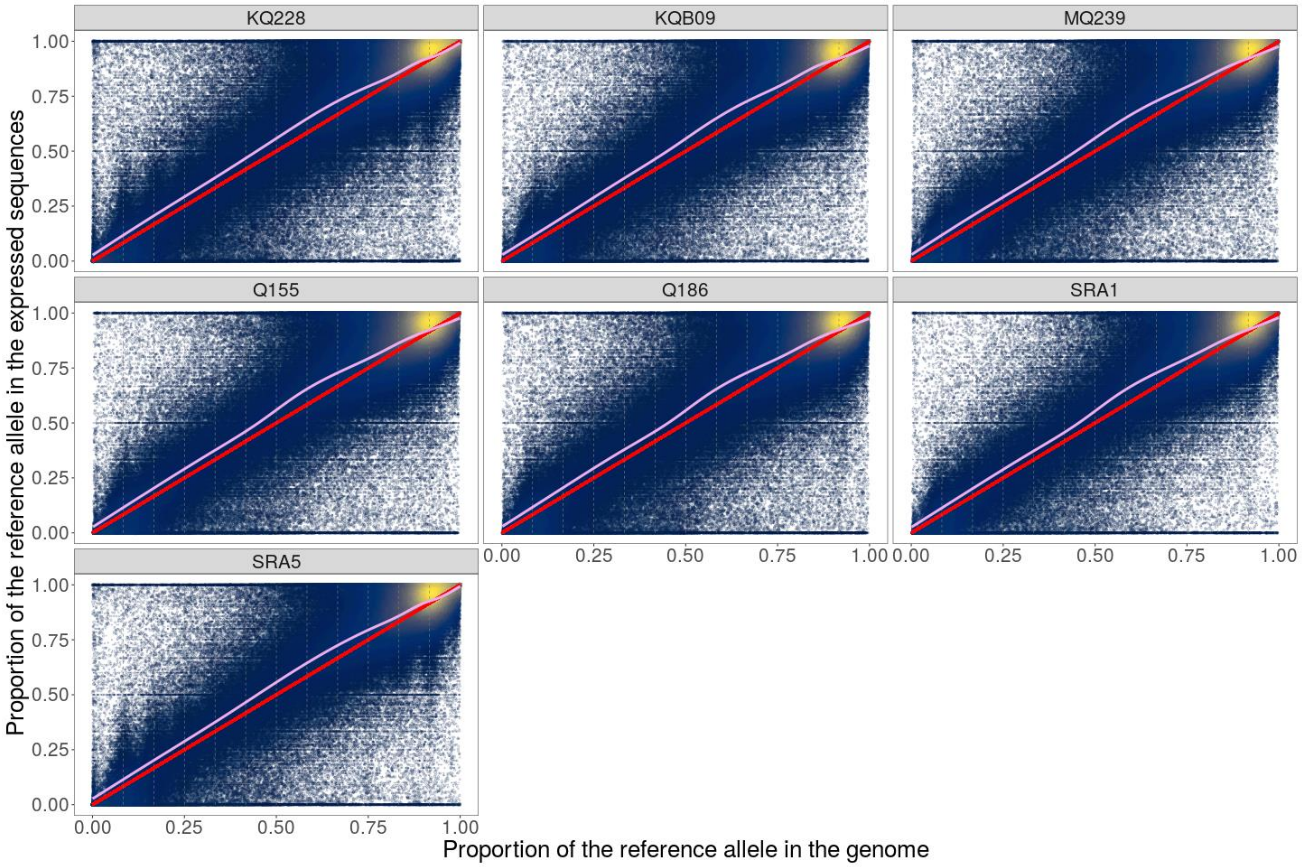
Supplemental Fig S2. Pedigree of the sugarcane hybrids used for allele-specific expression analyses. Arrows indicate parent → offspring relationships, and the size of the circle represents the number of crosses in which a genotype is involved. Genotypes often used in early sugarcane crossings are coloured in gold and Australian hybrids are coloured in green, with a darker tone for the seven hybrids sampled in our study. We highlight in purple the *Saccharum spontaneum* genotype Hainan92-9, grandparent of KQB09-20432.



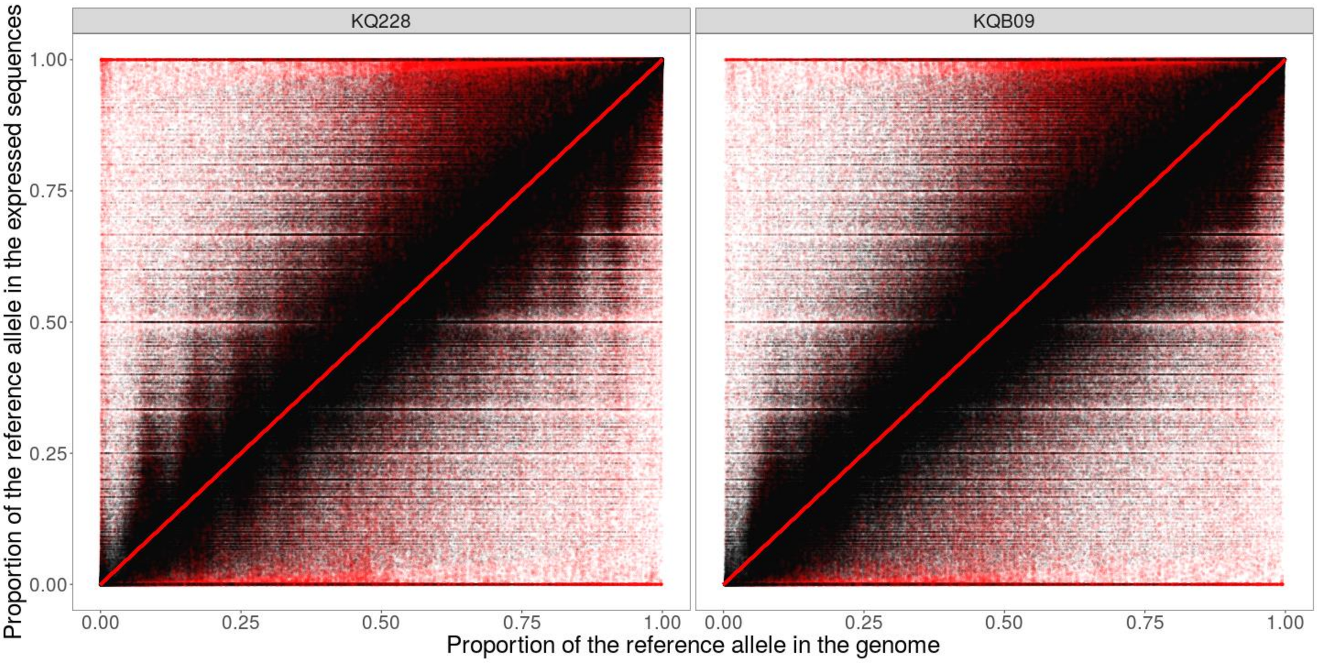
Supplemental Fig S3. Distribution of the median depth of coverage when aligning genomic reads from each genotype to the sorghum genome. A) Distribution for different gene features: CDS (coding sequence), exons and the entire gene (including introns). Note the peaks at roughly 1200 X or 240 X, which are more apparent for more conserved regions. B) Depth of coverage for exons, according to the classification of genes as single copy in sorghum, rice and Brachypodium. Only genes conserved in the three species were included.



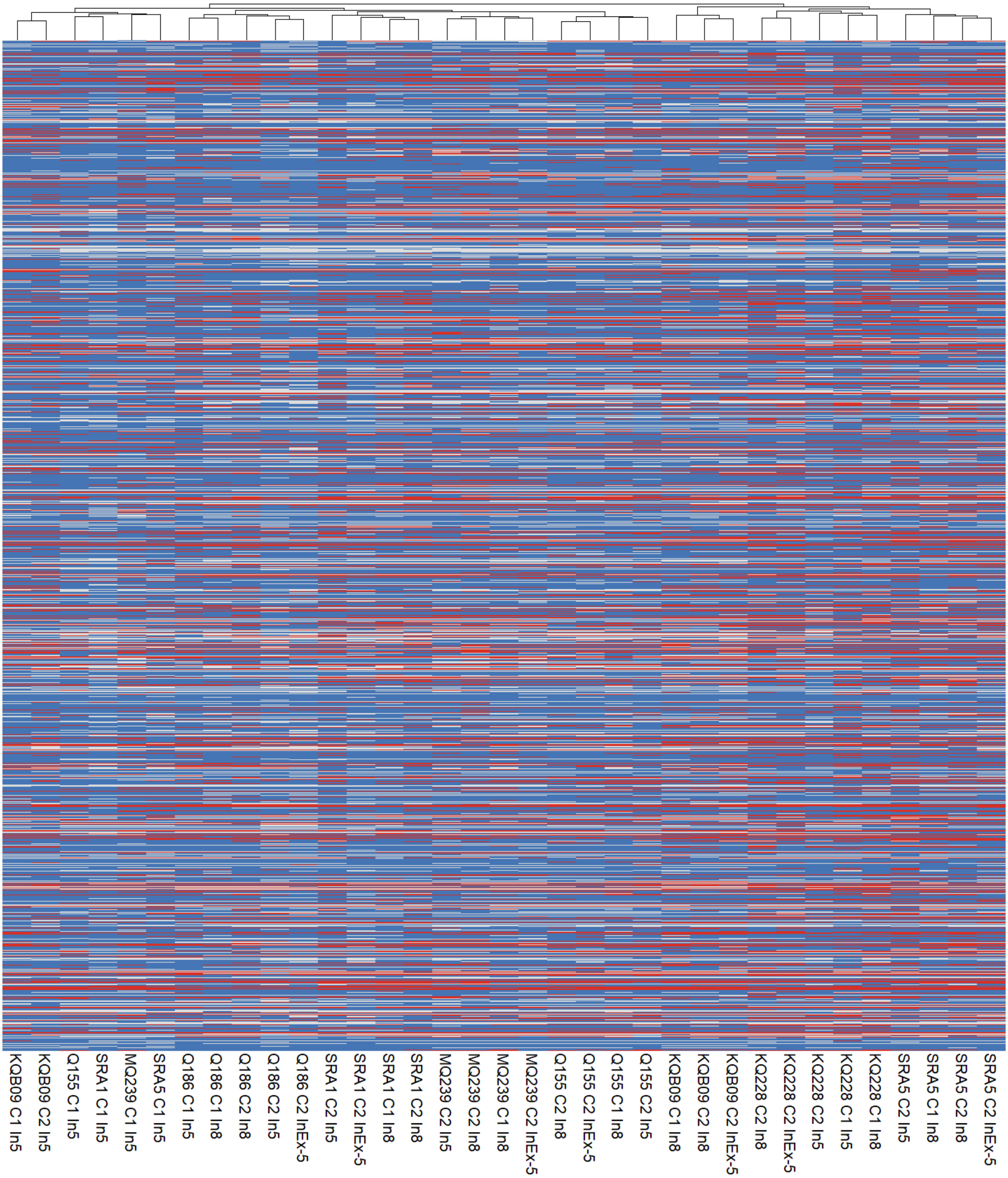
Supplemental Fig S4. Hierarchical clustering of samples based on the relative proportion of the reference allele. Euclidean distances were calculated considering heterozygous sites for which at least half of the samples had five or more RNA-seq reads. Seven clusters separate samples from each of the seven genotypes, with the exception of the first replicate of SRA1 C1 In5 (highlighted in red), which was grouped with MQ239 samples.



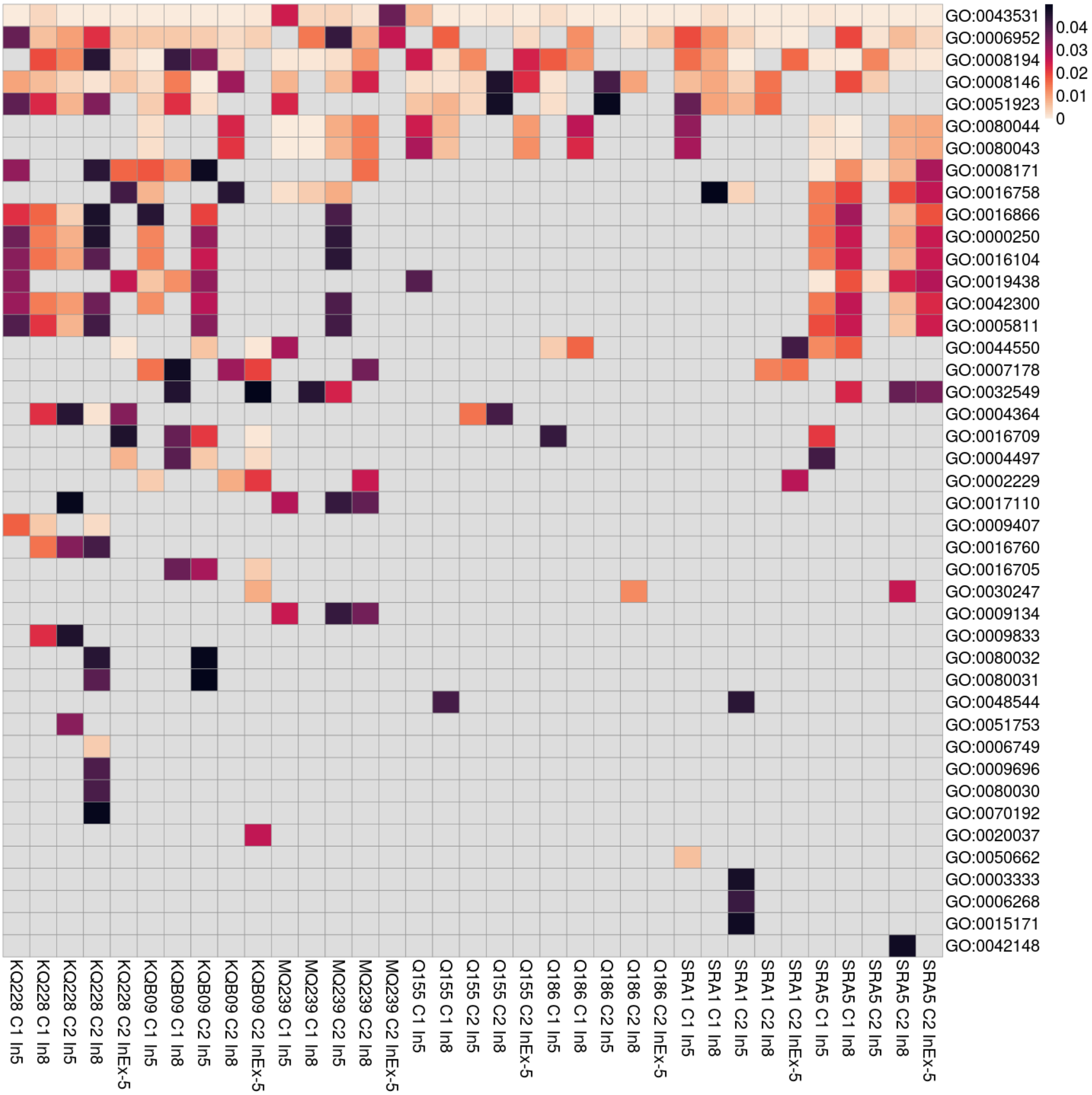
Supplemental Fig S5. Consistency of allele-specific expression across sugarcane genotypes. Each panel represents data for a different genotype, from internode 5 in the first collection. Note that clusters representing different genomic doses are less clearly visible because of the less stringent filters used (50 genomic reads and 10 RNA-seq reads).



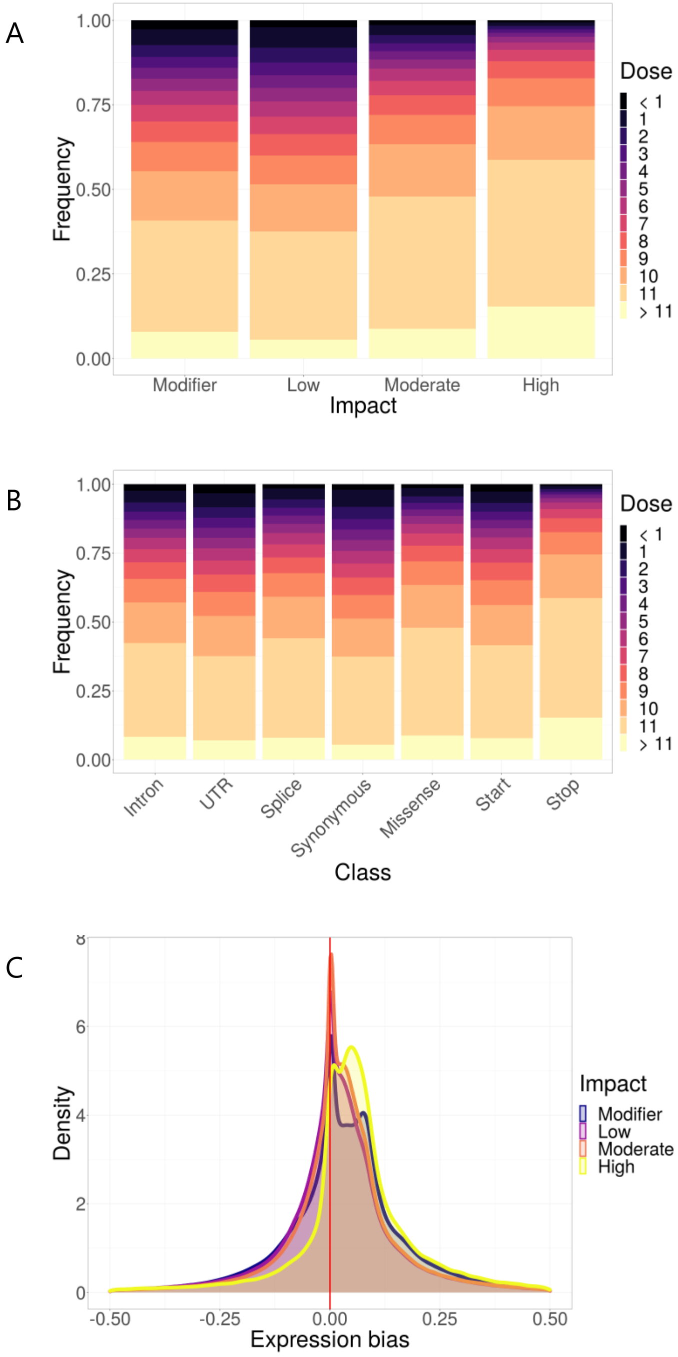
Supplemental Fig S6. Results of the allele-specific expression test. Each point represents one single nucleotide variant, and sites with significant ASE are indicated in red. We show the results for a genotype with high depth of coverage (KQ228) and for one with low sequencing depth (KQB09). Statistical power was comparable across the genotypes, partly because we modelled uncertainty both in the genomic and expressed allele ratios.



Supplemental Fig S7. Similarity of genotype × internode combinations based on the outcomes of allele-specific expression tests. Each row represents a gene, with blue indicating balanced allele expression and red indicating significant ASE. Note that most clusters include treatments from the same genotype, except for the immature internodes to the left.



Supplemental Fig S8. Functional enrichment of genes with allele-specific expression according to Gene Ontology terms. Significant false discovery rate-corrected p-values for the enrichment test are shown ($p < 0.05$), with lighter colours indicating lower p-values. Non-significant tests are shown in grey. For the full description of ontology terms see Supplemental Table S4.



Supplemental Fig S9. Comparison of single nucleotide variants with different predicted impact. A) Distribution of the genomic dose of the reference allele according to the predicted impact of each SNV. B) Distribution of the genomic dose of the reference allele for each class of mutation. Polymorphisms altering the stop codon showed the highest frequencies of the reference allele, followed by missense mutations and variants in splice sites and in start codons. C) Distribution of the deviation between RNA and DNA allele ratios (expression bias). Positive values indicate higher expression of the reference allele.