Limited allele-specific gene expression in highly polyploid sugarcane

Gabriel Rodrigues Alves Margarido (✉ gramarga@usp.br)

University of São Paulo <https://orcid.org/0000-0002-2327-0201>

Fernando Henrique Correr

University of São Paulo

Agnelo Furtado

University of Queensland

Frederik Botha

University of Queensland

Robert Henry

University of Queensland <https://orcid.org/0000-0002-4060-0292>

Article

Keywords: Saccharum, functional genomics, preferential allele expression

Posted Date: June 16th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-608681/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Genome Research on December 23rd, 2021.
See the published version at <https://doi.org/10.1101/gr.275904.121>.

Limited allele-specific gene expression in highly polyploid sugarcane

Running title: Limited allele-specific expression in sugarcane

Keywords: *Saccharum*, functional genomics, preferential allele expression

Gabriel Rodrigues Alves Margarido^{1,2*} – <https://orcid.org/0000-0002-2327-0201>

Fernando Henrique Correr^{1,2} – <https://orcid.org/0000-0001-8562-1946>

Agnelo Furtado² – <https://orcid.org/0000-0001-6130-9026>

Frederik C Botha² – <https://orcid.org/0000-0003-1082-6896>

Robert James Henry^{2†} – <https://orcid.org/0000-0002-4060-0292>

¹ Department of Genetics, University of São Paulo, "Luiz de Queiroz" College of Agriculture, Piracicaba 13418-900, Brazil

² Queensland Alliance for Agriculture and Food Innovation, University of Queensland, Brisbane 4072, Australia.

*gramarga@usp.br

†robert.henry@uq.edu.au

15 **Abstract**

16 Polyploidy is widespread in plants allowing the different copies of genes to be expressed
17 differently in a tissue specific or developmental specific way. This allele specific expression (ASE) has
18 been widely reported but the proportion and nature of genes showing this characteristic has not
19 been well defined. We now report an analysis of the frequency and patterns of ASE at the whole
20 genome level in the highly polyploid sugarcane genome. Very high-depth whole genome
21 sequencing and RNA-sequencing revealed strong correlations between allelic proportions in the
22 genome and in expressed sequences. This level of sequencing allowed discrimination of each of the
23 possible allele doses in this 12-ploid genome. Most genes were expressed in direct proportion to
24 the frequency of the allele in the genome with examples of polymorphisms being found with every
25 possible discrete level of dose from 1:11 for single copy alleles to 12:0 for monomorphic sites. The
26 rarer departures from allelic balance were more frequent in the expression of defence-response
27 genes, as well as in some processes related to the biosynthesis of cell walls. ASE was more common
28 in genes with variants that resulted in significant disruption of function. The low level of ASE may
29 reflect the recent origin of polyploid hybrid sugarcane. Much of the ASE present can be attributed to
30 strong selection in nature and domestication for resistance to diseases.

31 Introduction

32 Sugarcane has remarkable potential as a food and bioenergy crop. It is a C4 grass with high
33 photosynthetic efficiency and biomass yield ¹, grown mainly for the dual purpose of producing sugar
34 and ethanol. Currently, sugarcane comprises 86 % of the sugar crops cultivated worldwide ². In the
35 current scenario of rapidly changing global environmental conditions, biofuel obtained from
36 sugarcane can have a key role in addressing climate change concerns ^{3,4}. Breeding undertakings are
37 necessary to meet increasing demands for sugar and ethanol, and a better understanding of
38 molecular processes relevant to carbon partitioning will in turn guide molecular breeding of this
39 crop ⁵.

40 Genomic analyses in sugarcane (*Saccharum* spp.) are held back by the unique combination of
41 several layers of genomic complexity. All known *Saccharum* accessions are highly polyploid ⁶, with
42 evidence of both allo- and autopolyploidy events in the evolutionary history of the genus ⁷.
43 Commercially cultivated sugarcane clones are interspecific hybrids between *S. officinarum* and *S.*
44 *spontaneum*, with varying contributions from each genome, multiple aneuploidy events and
45 recombination between the parental genomes ⁸. Hybrids show a variable number of chromosome
46 copies per homology group ⁹, with different total numbers of chromosomes for distinct genotypes
47 ¹⁰, and have large genomes ¹¹. Still, there is a prevalence of $2n = 12x$ among interspecific hybrids
48 ^{12,13}. ¹⁴ have recently proposed the existence of three founding genomes in the genus *Saccharum*,
49 two of them unevenly found in the sugar-rich *S. officinarum*, and the third being observed in the
50 wild *S. spontaneum*. Last, polyploidy and vegetative propagation together promote high levels of
51 heterozygosity, and the combination of these factors affects breeding activities ¹⁵.

52 Various genomic resources are being developed for sugarcane ^{16,17}. However, because of this
53 amalgam of obstacles, researchers are still chasing the level of progress attained for other major
54 crops ¹⁸. One outstanding challenge is the assembly of a genomic sequence that fully represents a

55 complete hybrid genome. Despite the recent publication of multiple (partial) sugarcane genome
56 sequences ^{19–21}, sorghum is still a valuable genomic resource and has been extensively used as a
57 genome reference for sugarcane ^{9,22}. These two genomes are highly conserved, particularly in
58 transcribed regions ²³. The fact that sorghum is diploid means that using its genome sequence as a
59 reference sidesteps issues due to the multiple alignment of reads to several chromosome copies.

60 Polyploidy causes a multitude of changes in cell structure and function, both in allo- ^{24,25} and
61 autopolyploids ²⁶. It is usually accompanied by a so-called *genomic shock* and an increase in allelic
62 diversity, which can consequently cause dramatic changes to gene expression profiles ^{27–29}. Such
63 alterations include allele-specific expression (ASE), or preferential allele expression, a phenomenon
64 whereby the different alleles of a given gene are unevenly expressed ³⁰. This excessive abundance of
65 one allele stems from differential expression and/or degradation of mRNA molecules originated
66 from different chromosome copies. Allelic imbalance can occur due to variation in *cis*-acting
67 regulatory regions, nonsense-mediated decay and epigenetic imprinting, among other factors ^{31,32}.

68 ASE has been evaluated in diploid and polyploid plants, such as *Arabidopsis* ³³, maize ³⁴,
69 autotetraploid potato ³⁵ and allohexaploid wheat ³⁶. The joint availability of high-throughput RNA
70 sequencing assays (RNA-seq) and methods for quantitative genotyping of single nucleotide variants
71 (SNVs) has allowed ASE to be studied in sugarcane ^{37–40}. However, these previous investigations are
72 focused on a small number of genes, rely on limited genomic sampling or suffer from potential
73 sources of bias in the genotype calls. There are yet no genome-wide analyses of ASE in sugarcane
74 that make use of high-depth whole genome sequencing (WGS) data. Given the evidence of
75 pervasive ASE in hybrids ^{30,41}, the hybrid nature of sugarcane makes this an interesting system for a
76 more thorough exploration. Hybridization events in sugarcane include the likely occurrence of
77 allopolyploidy in its evolutionary history, interspecific hybridization during early breeding
78 endeavours, and crosses between elite genotypes to obtain full-sib progenies, which are all possible

causes of ASE. Artificial selection after interspecific hybridization, mainly for high sugar yield and disease resistance, has potentially affected the expression patterns of individual alleles in varying ways. Also, because many genes are involved in the accumulation of sucrose from top to bottom internodes in sugarcane^{42,43}, it is conceivable that ASE may play a role in this carbon partitioning process.

In this context, we aimed to gauge the genome-wide extent of ASE in culms of elite sugarcane hybrids. Using high-coverage resequencing data to accurately estimate the relative genomic proportion of alleles, we found that on average one out of every four genes showed significant preferential allele expression. We observed consistent patterns of ASE in sugarcane internodes collected at different developmental stages, and further uncovered associations between ASE and gene structure and function. We discuss how this approach increases our knowledge about gene expression in polyploid sugarcane and consider its consequences for breeding.

Results

We investigated the occurrence of ASE in a set of seven sugarcane hybrids, all representative of current elite breeding germplasm. According to the phenotypic characterization detailed in⁴⁴, KQ228 was chosen to represent a high yield (as measured in tonnes of cane per hectare), high early-season sugar genotype. In contrast, although SRA5 was also high-yielding, it showed low sugar and high fibre content. We additionally selected Q155 and KQB09-20432 (denoted herein by KQB09) to represent high sugar and high fibre genotypes, respectively. SRA1 showed both low sugar and low fibre content in this trial and was also used in our work. Finally, we chose MQ239 and Q186 due to their intermediate behaviour with regard to both traits.

To identify polymorphic sites we first sequenced the entire genomes of the seven hybrids. We sequenced KQ228 and SRA5 at a depth of coverage of 100 X per chromosome copy, or 1,200 X

102 of the monoploid genome, considering a polyploid genome of 10 Gbp with 12 copies of each
 103 chromosome ^{11,12}. For the other genotypes we obtained 240 X of the monoploid genome
 104 (Supplementary Table 1). We aligned the high quality reads against the sorghum genome reference
 105 and achieved alignment rates varying between 37.10 % and 38.48 %. Specificity of alignment against
 106 genic regions was high, with little effect of filtering out low quality alignments, in agreement with
 107 the fact that these are low-copy, conserved regions. As an example for KQ228, we note that the
 108 effective depth of coverage after alignment was close to the expected value of 1,200 X
 109 (Supplementary Fig. 1A). In particular for conserved single copy grass genes, which are more likely to
 110 also be present in sugarcane, we noticed that our approach was successful in unambiguously
 111 assigning reads to genes (Supplementary Fig. 1B).

112 We identified a total of 9,321,181 polymorphic sites in the genic regions of the seven
 113 genotypes, of which 6,533,916 were biallelic SNVs. The majority of the remaining sites were
 114 insertions and deletions, indicating that the redundancy brought about by the multiple chromosome
 115 copies in sugarcane possibly allows for a substantial amount of indels to be present. Next we filtered
 116 out low quality polymorphisms and kept 5,748,251 biallelic SNVs. By classifying these sites based on
 117 their predicted impact on gene structures, we found the following distribution of predicted effects:
 118 13,619,337 (79.82 %) modifier effects; 1,799,543 (10.55 %) SNVs with low impact; 1,608,014 (9.42 %)
 119 with moderate effects; and 36,392 (0.21 %) with high predicted impact. Note that the number of
 120 effects is larger than the number of SNVs, because each site can be annotated multiple times when
 121 in close proximity to multiple genes. For SNVs present within coding regions, 1,567,565 (48.96 %)
 122 were annotated as silent, or synonymous substitutions; 1,611,587 (50.33 %) caused amino acid
 123 changes (missense substitutions); and 22,716 (0.71 %) were nonsense mutations. It is also interesting
 124 to note that 2,140,328 of the 5.75 million high quality SNVs were polymorphic only in comparison to

the sorghum reference, but fixed in these sugarcane genotypes. Because we can only assess ASE for heterozygous sites, these loci were not used for downstream analyses.

The expression dataset we used corresponds to an RNA-seq assay of sugarcane stalks at five different developmental stages. For each of the genotypes we had triplicate libraries for internodes of different ages (collection time points C1 and C2) and varying maturity (internodes 5, 8 and 22, the latter also denoted by Ex-5 – see the Methods section for details). We had from 56.02 to 85.11 million raw reads available per library, of which 50.57 % to 84.67 % remained after preprocessing to remove low quality reads and contaminant ribosomal RNA, with a median yield of 81.79 % (Supplementary Table 2). We also used hierarchical clustering to assess the variation between biological replicates and thus to diagnose potential issues during library prep and sequencing (Supplementary Fig. 2). With these sources of information we finally discarded four samples of lower quality: one biological replicate each of KQ228 C2 In5, SRA1 C1 In5, Q186 C2 In5 and Q186 C2 InEx-5.

As expected, alignment rates of the RNA-seq libraries against sorghum were higher than for the WGS data, ranging from 72.41 % to 85.98 % (Supplementary Table 2). After quantifying allele-specific abundances for each biallelic SNV, we filtered out poorly covered genomic regions and lowly expressed genes and analysed from 512,827 to 851,295 heterozygous sites per treatment level (Table 1).

We initially carried out an exploratory data analysis to investigate the relationship between genomic and expressed proportions of the reference allele at each site. By doing so several features of ASE in sugarcane became readily apparent. First, the genomic allele proportions revealed eleven well-defined clusters, with each cluster centred directly over a fraction of twelve (Fig. 1). This agrees with there being eleven heterozygous classes, with one to eleven doses of the reference allele in a dodecaploid genome. We also noticed that the majority of SNVs showed eleven copies of the major

allele, i.e., the alternative allele was present in a single dose for most variant sites. Some loci did not cluster tightly with any of the groups, including several with apparent dose greater than eleven or less than one. In addition to simple random noise, this may indicate events of aneuploidy and/or copy number variants.

Overall we observed strong agreement between genomic and expressed allele ratios. The fraction of reads carrying the reference allele in the RNA-seq dataset was directly proportional to the corresponding WGS fraction (Fig. 1). Also, we saw very limited bias towards the reference allele, which is a common concern in ASE studies³¹. Another interesting feature we found is that many SNVs showed exclusive expression of one allele, as seen by the masses of points forming the horizontal lines at $Y = 0$ and $Y = 1$ in that figure. In fact, these loci are the likely cause of the departure of the smoothed trend (in pink) from the null expectation (red diagonal line in Fig. 1). We stress that these observations were consistent for all genotypes and internodes (Supplementary Fig. 3).

A Bayesian hierarchical Beta-Binomial model³⁷ confirmed that most SNVs showed no evidence of ASE, with 19,154 to 56,956 heterozygous sites yielding significant tests (Table 1 and Fig. 2). We noticed similar fractions of SNVs with ASE for all genotypes and internodes, except for the four cases where a lower quality sample was discarded. Uniformity was also seen when we gathered the heterozygous loci at the gene level, as the number of gene models with at least one informative SNV ranged from 16,538 to 19,254. Of these, between 2,311 and 5,902 had at least two SNVs with significant test results, and were classified as genes showing allele-specific expression. They correspond to 13.65 % to 31.71 % of the effectively sampled genes, with an average of 24.09 % (Table 1). The full set of genes with ASE, for the 35 combinations of genotypes and internodes, is provided in Supplementary Table 3.

172 When comparing the lists of genes showing ASE for the various treatments we found higher
 173 similarity among samples from the same genotype than from the same internode (Fig. 3). On
 174 average, the percent overlap of genes with ASE for different internodes of the same genotype was
 175 61.58 %, ranging from 57.75 % (SRA1) to 63.90 % (KQ228). In some cases the overlap was as high as
 176 74.60 %, as seen for the pair KQ228 C2 In8 and KQ228 C2 InEx-5. On the other hand, the overlap
 177 among samples of the same internode ranged from 52.34 % (C1 In5) to 57.46 % (C2 In8), with an
 178 average of 54.76 %. Interestingly, we also found that combinations involving different genotypes
 179 and different internodes still showed noticeable overlap, ranging from 31.80 % to 61.69 % (average
 180 of 49.50 %). We therefore note that some genes consistently showed ASE for different treatments,
 181 while there was an added effect of internode and an even greater contribution of the genotype
 182 factor.

183 A more detailed comparison of heterozygous SNVs in common for KQ228 and SRA5 showed
 184 a strong positive correlation between the genomic allele proportions observed in each genotype
 185 ($\rho = 0.92$, with $p\text{-value} < 10^{-15}$) (Fig. 4A). Rarely was the difference between their most likely doses
 186 greater than three or four chromosome copies. As a consequence, we saw a similar trend for the
 187 relationship between the expressed allele proportions, with $\rho = 0.88$, $p\text{-value} < 10^{-15}$ (Fig. 4B). In
 188 any case, we did observe some loci for which there were substantial differences between the
 189 proportions of the reference allele in the two transcriptomes.

190 Even more noteworthy, when there was a significant excess of a given allele in both
 191 genotypes, we found that the preferential allele expression often occurred in the same direction (Fig.
 192 4C). In most cases the bias was towards the reference allele. These observations are likely reflections
 193 of their similar genetic background. This may be in part due to the shared domestication history of
 194 *Saccharum* genotypes, as well as to the recent breeding of sugarcane. All genotypes investigated

195 here represent elite germplasm used in breeding programs, and so have been through strong
 196 selective pressures.

197 We were interested in understanding whether the occurrence of ASE was more frequent in
 198 genes involved in particular cellular functions. To that end we applied the Gene Set Enrichment
 199 Analysis (GSEA) strategy ^{45,46} and found evidence of enriched Gene Ontology (GO) terms ^{47,48}. Two
 200 functional terms appeared as consistently enriched for most treatments, namely *ADP binding*
 201 (GO:0043531) and *defence response* (GO:0006952) (Supplementary Table 4 and Supplementary Fig.
 202 4). These terms are in fact closely related, as many resistance (R) proteins have nucleotide-binding
 203 domains, and many genes were simultaneously annotated with both terms. Interestingly, we also
 204 found enrichment of *UDP-glycosyltransferase activity* (GO:0008194) for the majority of genotypes
 205 and internodes. UDP-glycosyltransferases are a large protein family with multiple roles, including the
 206 biosynthesis of many cell wall compounds. The terms *sulfotransferase activity* (GO:0008146) and
 207 *sulfation* (GO:0051923) were also enriched for many treatments. Additional GO terms were
 208 significantly enriched in particular treatments, but no discernible pattern was observable, except for
 209 an apparent excess of terms for KQ228 and SRA5, the genotypes sequenced at higher depth. It is
 210 noteworthy that we found the terms *O-methyltransferase* (GO:0008171) and *aromatic compound*
 211 *biosynthetic process* (GO:0019438) enriched in the fibre-rich genotype SRA5, particularly in its upper
 212 internodes. Manual inspection revealed that genes annotated with these terms and showing ASE are
 213 potentially involved in the synthesis of cell wall components. For instance, they include a gene
 214 similar to ZRP4 (*Zea Root Preferential*), which is involved in the synthesis of suberin and possibly of
 215 lignin ^{49,50}. We also saw evidence of strong ASE for a gene similar to Herbicide safener binding
 216 protein, an O-methyltransferase predicted to be involved in the synthesis of lignin precursors ⁵¹. We
 217 further noticed that many of the SNVs in genes annotated with these enriched terms showed
 218 exclusive expression of the reference allele (Fig. 5A and 5B). This extreme form of ASE may imply

219 that transcriptional and post-transcriptional regulatory mechanisms have a potentially strong impact
220 on the abundance of different alleles for these genes.

221 A comparable tendency was seen for SNVs predicted to have a large impact on protein
222 structure (Fig. 5C). This classification is assigned to SNVs that affect the stop codon, possibly giving
223 rise to truncated or extended proteins, in the case of stop gain and stop loss mutations, respectively.
224 Such gene products are often not fully functional and can be detrimental to cell activity. It is thus
225 not surprising that we observed a single allele to be exclusively present for many of these loci,
226 providing evidence of absence of transcription or post-transcriptional mRNA decay of the other SNV
227 allele⁵². At the genome level, the distribution of estimated allele doses per SNV class showed an
228 increase in the frequency of higher doses of the reference allele for variants of higher predicted
229 impact (Supplementary Fig. 5A). Polymorphisms altering the stop codon showed the highest
230 frequencies of the reference allele, followed by missense mutations and variants in splice sites and in
231 start codons (Supplementary Fig. 5B). In addition to the already higher genomic doses, we also
232 observed a higher than expected abundance of transcripts carrying the reference allele for high-
233 impact SNVs (Supplementary Fig. 5C).

234³⁵ found that core angiosperm genes were more likely to show ASE in multiple tetraploid
235 potato genotypes. Similarly, we found that genes conserved in the monocots had higher odds (>
236 1.5) of showing ASE as the other genes (Fig. 5D). Conversely, genes that were found to be single
237 copy in sorghum, rice and Brachypodium were less likely to show preferential expression of alleles.
238 These observations corroborate the notion that genes controlling key biological processes may be
239 enriched among those with ASE, while suggesting that single copy genes may be under some
240 control mechanism inhibiting ASE in sugarcane. A trend of there being lower odds of ASE for
241 sorghum-exclusive paralogs was also noticeable, but their number was limited and statistical
242 significance was often not attained (Fig. 5D).

243 Discussion

244 We report limited levels of allele-specific expression in the culms of a set of elite sugarcane
245 hybrids. Most heterozygous loci showed agreement of allele ratios in genomic and expressed
246 sequencing reads, with a clear linear relationship between these two variables. This suggests that
247 mechanisms regulating gene expression have comparable effects on the alleles of a large fraction of
248 the genes. Previous studies had shown localized evidence of ASE for particular sugarcane genes^{38–40},
249 but our work represents the first genome-wide analysis of high-depth whole-genome sequencing
250 data in this crop. This comprehensive dataset revealed that ASE may not be a widespread
251 phenomenon in the highly polyploid and complex sugarcane genomes.

252 The sequencing design we employed allowed us to assess ASE with higher accuracy in
253 KQ228 and SRA5, while still achieving high depths to compare all seven genotypes. Due to the high
254 ploidy levels of sugarcane hybrids, it is essential to make use of high sequencing depths to obtain
255 precise estimates of allele doses^{53,54}. This is particularly relevant if one considers the possibility of
256 bias and overdispersion in allele read counts. Also, to err on the side of being conservative we
257 removed four RNA-seq libraries from all ASE analyses. We opted for this more conservative
258 approach to avoid potential issues in accurately quantifying the expression of different alleles.
259 Nevertheless, we note that these samples were from distinct treatments, such that the effect on
260 statistical power to detect ASE was limited.

261 Because we performed cross-species read alignment, we included in our pipeline various
262 filtering steps to reduce the occurrence of spurious SNVs, while also allowing a large fraction of
263 reads to be used. In any case, alignment of reads from transcribed regions is relatively
264 straightforward, such that the majority of relevant alignments were unambiguous. Ensuring that a
265 large proportion of the reads are effectively and correctly assigned to their corresponding genomic
266 regions is crucial in ASE studies³¹. The fact that we observed very little (if any) bias towards the

reference allele indicates a good overall performance of the pipeline, underlining the reliability of our results.

Variations of a Beta-Binomial model are frequently used for investigating ASE in a variety of scenarios^{55,56}. Such a model was adapted for mixed- and high-ploidy organisms³⁷, and we further modified this method to model uncertainty in estimating allele doses. This approach allows non-homogeneous sequencing depth profiles to be naturally handled by the model, with a corresponding adjustment of the null hypothesis for each evaluated polymorphism. Our statistical analysis strategy likely had an impact on the fraction of genes identified as showing significant ASE. While we found evidence of ASE for 13.65 % to 31.71 % of the genes, other researchers observed widely varying values: 13.5 % in *Arabidopsis*³³, roughly a third to half of the genes in potato³⁵, ~ 50 % of the genes in maize³⁴, and up to 70 % among wheat homoeologs³⁶. This variation encompasses differences between diploids, auto- and allopolyploids, as well as true biological variation in these various systems. It also includes differences in experimental designs and environmental and technical noises. Still, we cannot rule out a contribution from the distinct data processing and methodologies used in each case. Different statistical models and testing strategies, hard thresholds (e.g., for the minimum number of SNVs with ASE per gene) and other *ad hoc* criteria can affect the lists of significant genes. In sugarcane,³⁷ also observed ASE for approximately 40 % of the sampled genes. However, this study relied on genotyping-by-sequencing of reduced representation libraries, a strategy that is prone to biases and genotyping errors⁵⁷. These issues can then strongly influence conclusions about ASE for a set of the polymorphic sites.

In this context, one advantage of the GSEA method we used is that it does not depend on previous testing for ASE, but only on the ranking of genes. Genes with large absolute deviations between genomic and expressed allele ratios had higher ranks, and functional terms with many high-ranking genes were tagged as enriched with allele-specific expression. A small number of GO

291 terms showed significant enrichment, with the prominent presence of defence-response genes
 292 among those with ASE in many genotype and internode combinations. Many resistance proteins
 293 have nucleotide-binding domains, and binding to ADP or ATP induces conformational changes that
 294 regulate signalling cascades ^{58,59}. Hence, the large number of genes jointly annotated with defence
 295 response and ADP-binding terms explains the two major GO terms we detected. The higher
 296 frequency of ASE among defence-response genes agrees with observations in allohexaploid wheat,
 297 where many of these genes showed homoeolog expression bias in plants challenged with a
 298 pathogenic fungus ³⁶. Species in the genus *Saccharum* differ broadly with regard to their response to
 299 pathogen infection, and sugarcane hybridization and breeding have historically focused on selecting
 300 for disease resistance ^{60,61}. Given the strong selective pressure on defence-response genes, it is not
 301 surprising that many of them showed evidence of ASE. Different alleles may respond differently to
 302 particular pathogen strains or races, and breeders may have indirectly selected for higher expression
 303 of specific alleles. This ASE probably is also linked to the smaller contribution of the *S. spontaneum*
 304 genome to modern hybrids, as a consequence of abnormal chromosome transmission in the initial
 305 backcrossing. Selection pressure on disease resistance kept those traits from *S. spontaneum* in the
 306 genome and for most they are left in low doses.

307 We did not detect homogeneous enrichment of many functional terms associated with
 308 carbon partitioning, such as those involved in sucrose accumulation or cell wall biosynthesis, with
 309 the exception of UDP-glycosyltransferase activity. This does not mean, however, that ASE has no
 310 substantial contribution to these biological processes. We observed enrichment of assorted
 311 transferase terms (including of hexosyl groups), *cellulose synthase (UDP-forming) activity*
 312 (GO:0016760), *plant-type primary cell wall biogenesis* (GO:0009833) and *mannan synthase activity*
 313 (GO:0051753) for scattered treatments. This indicates that sizeable ASE may occur at least in some
 314 genotypes and at some points during the development and maturation of sugarcane stalks. Also,

the identification of high-level GO terms associated with the synthesis of cell wall compounds in SRA5 opens up the possibility that ASE has direct effects on phenotypes of economic interest.

When analysing high-impact SNVs, which are predicted to have more drastic effects on protein structure, we discovered both higher genomic doses and excessive abundance of the reference allele. The former implies purifying selection against the alternative allele, while the latter hints at its silencing or increased mRNA degradation. This effect of the type of mutation on the incidence of ASE has been reported in humans, similarly with a higher frequency of ASE in nonsense variants⁵². Indeed, polymorphisms affecting start and stop codons may be the primary cause of ASE for many genes. If this is in fact the case, we would expect the genes bearing these SNVs to show similar expression patterns across tissues and developmental stages. This agrees with our observation that the transcriptomes of different internodes for the same genotype showed more similar patterns of ASE than other combinations of treatments (Fig. 3). Hence, the control of ASE in sugarcane seems to be largely genetic, which has important implications for plant breeding. The expression of favorable alleles can be directly or indirectly leveraged through selection, at least for a subset of the genes, and these effects are expected to be partly passed onto the progeny of crosses between hybrids. Yet, although genotype was the main factor determining the occurrence of ASE, we did also observe an effect attributable to internodes and collection time points. In other words, some genes with significant ASE were found in common for the same collection and/or internode of different genotypes. Developmental and environmental cues thus appeared to have some contribution to ASE in our experiment. These effects are also partially accessible to selective breeding, to the extent that they may contribute to genotype × internode interaction. They are also interesting from a biological perspective, as they enhance our understanding of the biology of sugarcane hybrids.

Clusters of genes conserved in the Liliopsida were overrepresented among those with ASE, while single copy grass genes were underrepresented. Our results support those obtained in autotetraploid potato, where genes responsible for key plant processes have been shown to be enriched in ASE³⁵. Fine-tuning of biological processes that are essential for proper functioning of plant metabolism may include mechanisms favouring the expression of some alleles over the others. On the other hand, the lower frequency of ASE among single copy genes suggests that redundancy may be a relevant factor driving ASE in sugarcane. We hypothesize that the extended mutational freedom afforded by gene duplication allows regulatory mechanisms to tweak the expression of individual allelic copies in these high-ploidy genomes.

The genomic data we analysed supports the presence of 12 chromosome copies for the majority of polymorphic sites in these seven sugarcane hybrids, corroborating recent findings¹³. Despite the extensive variation in ploidy and chromosome number existent in *Saccharum*, it is compelling to see such consistency among a set of elite hybrids. As is common in sugarcane breeding programs, these hybrids share a narrow genetic base and are often derived from crosses involving repeated parents. We can presume that their common genetic background may have increased the likelihood of producing genomes with similar composition, in spite of their substantial phenotypic variation. Our high-depth whole-genome sequencing strategy was paramount in verifying that $2n = \sim 12x$ for most genomic regions, which include a mixture of hom(oe)ologs of varying origin. However, we could neither infer the genomic origin of each allele nor observe any clear differentiation between putative *S. officinarum*- and *S. spontaneum*-derived alleles based on short-range polymorphisms alone. Although the genus *Saccharum* may comprise three distinct subgenomes, these founding genomes are apparently very similar to each other, especially for exonic regions, where the average sequence identity is greater than 99 %¹⁴. This observation agrees with²¹, who found that coding sequences of hom(oe)ologous genes in a sugarcane hybrid have a

362 median sequence divergence of less than 1 %. Given such high similarity among gene copies, we
 363 should indeed not expect to be able to tell the three subgenomes apart when working at the SNV
 364 level in transcribed regions. Hence, our results agree with the observations of ¹⁴ and further suggest
 365 that this high sequence similarity likely extends across most of the sugarcane genes.

366 The combination of high-depth WGS data and RNA-seq has provided a comprehensive view
 367 of the incidence of allele-specific expression in sugarcane hybrids. It does not appear to be a
 368 pervasive feature of gene expression in this complex crop. By using the sorghum genome as a
 369 reference we have focused on genes that are conserved between the two species. An interesting
 370 future research opportunity is to investigate sugarcane-exclusive genes and assess whether they are
 371 less likely to show ASE. This will be possible when we have a complete (phased) hybrid sugarcane
 372 genome available. In addition, this genomic resource will allow us to explore polymorphisms in
 373 regulatory regions, and consequently to identify *cis*-acting elements that control the occurrence of
 374 ASE in sugarcane. Then, we can leverage this information to lead molecular breeding efforts in this
 375 important bioenergy crop.

376 **Methods**

377 *Biological material and sample collection*

378 ⁴⁴ describe a field experiment used to evaluate 24 sugarcane hybrids contrasting in several
 379 traits, including fibre content and sugar yield. Later, we carried out RNA-seq to assess differential
 380 gene expression in internodes collected at five different developmental stages. Briefly, 24 *Saccharum*
 381 hybrids were planted in a 6 × 4 Latin square with three replicates, in August 2017 in Queensland,
 382 Australia. These genotypes were phenotypically characterized in March, June and September 2018,
 383 by measuring the soluble solids content, polarity and fibre percentage. For RNA-seq, three
 384 independent biological replicates of stalks were collected 19 weeks (the first collection, or C1) and

385 37 weeks (C2) after planting. Internodes 5 and 8 were sampled in both collections, while internode
 386 Ex-5 was sampled in C2. The latter corresponds to internode 5 from C1, tagged in intact culms to be
 387 collected when more mature. We refer the reader to ⁴⁴ for further details about the trial.

388 Based on these phenotypic traits we chose seven genotypes to use in our work. KQ228 and
 389 SRA5 showed contrasting behaviour in terms of sugar and fibre accumulation, and were chosen as
 390 the main genotypes to explore herein. We also sampled the pair Q155 and KQB09-20432, which
 391 showed similar phenotypes to KQ228 and SRA5, respectively. The remaining genotypes, SRA1,
 392 MQ239 and Q186, were selected because of their negative or neutral phenotypic traits.

393 *Whole genome sequencing and quality control*

394 We took samples from leaves +4 and +5 of each of the seven genotypes for whole genome
 395 sequencing, where leaf +1 is the first with a visible dewlap. DNA was extracted using the method
 396 described by ⁶², modified by adding 5 ml chloroform instead of the phenol:chloroform:isoamyl
 397 alcohol (25:24:1) mixture. Following extraction, DNA integrity was checked with a NanoDrop
 398 spectrophotometer assay to measure the 260/280 and 230/260 ratios.

399 High-quality genomic DNA was used to build sequencing libraries following the Illumina
 400 protocol. The seven libraries were sequenced on one S4 flowcell of the NovaSeq 6000 platform to
 401 obtain 2 x 150 bp paired-end reads. We further designed the sequencing run to represent the
 402 genotypes at different depths of coverage. Libraries for KQ228 and SRA5 were sequenced at five
 403 times the volume of the other genotypes, with expected depths of 100 X and 20 X per chromosome
 404 copy, respectively.

405 Raw whole-genome reads were trimmed using CLC GENOMICS WORKBENCH V20.0.4 at a quality
 406 limit of 0.01 to remove data with low Phred scores (Supplementary Table 1). We also removed reads
 407 with more than two ambiguous bases and discarded reads shorter than 50 bases after trimming.

408 *Read alignment and SNV calling*

409 We used the *Sorghum bicolor* genome 454 v3.0.1^{63,64} as a reference to identify SNVs. To the
 410 original genome sequence and annotation, downloaded from Phytozome v13⁶⁵, we collated
 411 annotations from Ensembl Plants 49⁶⁶ using BIOMART v2.44.4^{67,68} in R v4.0.5⁶⁹.

412 Alignment of genomic reads to the sorghum genome was performed with BOWTIE2 v2.4.1⁷⁰.
 413 We used end to end (global) alignment with the very sensitive setting, while requiring fragment sizes
 414 to be between 50 and 800 bp. We further modified the minimum score function (--SCORE-MIN L,0.0,-
 415 0.8) and indel penalty (--RDG 4,1 --RFG 4,1) to deal with cross-species alignment. After read sorting
 416 and ordering with SAMTOOLS v1.10⁷¹, we marked duplicate reads with PICARD TOOLS v2.23.4
 417 (<http://broadinstitute.github.io/picard>). Next we filtered reads to remove secondary and
 418 supplementary alignments, as well as PCR and optical duplicates. Only reads with a minimum
 419 mapping quality of three were kept, as this allows for a limited number of mismatches between the
 420 read and reference sequences.

421 We used the GATK v4.1.8.1 pipeline⁷² to identify SNVs and call genotypes. For doing so we
 422 first ran HAPLOTYPECALLER individually for each sample, restricting the genomic intervals to annotated
 423 genes and with a padding of 100 bp on both sides. We set the ploidy to 12, with a heterozygosity
 424 value of 0.01 and indel-heterozygosity of 0.001. Next we ran GENOMICSDBIMPORT and
 425 GENOTYPEGVCFs to identify variants and initially filtered the sites to keep only biallelic SNVs. We then
 426 loaded the SNVs with VARIANTANNOTATION v1.36.0⁷³ and filtered sites according to the following
 427 criteria: total depth between 100 and 10,000, quality score ≥ 50 , FISHERSTRAND ≤ 80 and $-8 <$
 428 READPOSRANKSUM < 8 . We also removed monomorphic sites, which were different from the sorghum
 429 reference, but fixed in our sugarcane samples. Last, we annotated the predicted impact of each SNV
 430 based on their relative genomic position with SNPEFF v5.0⁷⁴.

431 *RNA sequencing, data preprocessing and alignment*

432 Total RNA was extracted from three replicates of each genotype × internode combination.
 433 We used a total of 105 samples representing 35 treatments (7 genotypes × 5 internodes), with each
 434 sample comprising a pool of four clonal stalks.

435 Sequencing libraries were constructed with the standard TruSeq RNA protocol and the cDNA
 436 was sequenced (in pools with additional indexed libraries) on the Illumina NovaSeq 6000. Samples
 437 were split into four lanes of one S4 flowcell, with a data volume corresponding to 90 samples per
 438 lane. With this multiplexing strategy the expected yield was of more than 50 million paired-end 100
 439 bp reads for each sample (Supplementary Table 2).

440 We first preprocessed the raw reads with TRIMMOMATIC v0.39⁷⁵ to remove Illumina adapters,
 441 allowing a maximum of two mismatches and using parameters PALINDROMECLIPTHRESHOLD 30 and
 442 SIMPLECLIPTHRESHOLD 10. The minimum Phred quality for leading and trailing bases was set to 3, and
 443 a sliding window of size four, with an average required quality score of 30, was used to remove low
 444 quality bases. We also removed the 13 leading bases to account for random hexamer priming bias.
 445 After trimming we removed reads shorter than 70 bp. Next we used BBDOUK v38.79
 446 (sourceforge.net/projects/bbmap) to remove contaminating ribosomal RNA reads, based on the
 447 SILVA rRNA database⁷⁶ and a kmer length of 31.

448 For aligning the RNA-seq reads we used the same sorghum genome reference as for the
 449 WGS data. To account for introns in the genome we used the splice-aware aligner HISAT v2.1.0⁷⁷. As
 450 parameters we used the known splice sites from the structural gene annotation, with a maximum
 451 intron length of 20 Kbp, and again modified the minimum score function (--SCORE-MIN L,0.0,-0.8) and
 452 indel penalty (--RDG 4,1 --RFG 4,1).

453 *Quantification of relative allele proportions*

454 For both the WGS and RNA-seq datasets we applied the ASEReadCounter tool v4.1.8.1 to
 455 estimate the frequency of each allele, in read counts, for each polymorphic site of each genotype. At
 456 the genomic level, this provided an estimate of the relative allele dose. In this case we removed
 457 PCR/optical duplicates to avoid potential bias in estimating the doses. At the expression level, we
 458 used these counts to obtain estimates of the relative expression levels of both alleles. For RNA-seq
 459 reads we disabled the duplicate read filter, so as not to bias the estimated allelic proportions for
 460 highly expressed genes. In both cases we disabled the maximum depth limit to ensure that all reads
 461 were effectively counted. As a diagnostic metric we performed hierarchical clustering with R v4.0.5 to
 462 visualize the dissimilarity among samples. To that end we used SNVs for which the RNA-seq read
 463 count was greater than or equal to five for at least half of the samples. Euclidean distances were
 464 calculated based on the relative frequency of the reference allele and the default complete linkage
 465 method was used to find clusters.

466 *Assessment of allele-specific expression*

467 Based on the clustering pattern and the amount of rRNA and PCR duplicates, we excluded
 468 four samples that were of comparatively lower quality than the rest in the RNA-seq dataset. These
 469 samples were all from distinct treatments, such that four genotype $\times$ internode combinations were
 470 represented by two replicates, while the remaining 31 treatments had all three replicates. For each of
 471 the 35 treatments we combined the allele counts of the remaining high quality RNA-seq replicates.

472 In addition to the raw allele counts, we calculated for each polymorphic site the relative
 473 proportion of the reference allele, both for the genomic and expression datasets. Let r_{ik} represent
 474 the number of WGS reads carrying the reference allele for SNV i and treatment k , and let g_{ik}
 475 indicate the total number of genomic reads for the corresponding locus. We then calculated the

476 observed *genomic* proportion of the reference allele, denoted by Γ , as $\Gamma_{ik} = \frac{r_{ik}}{g_{ik}}$. Similarly for the
 477 transcriptome data, we denote by y_{ik} the number of RNA-seq reads with the reference allele and by
 478 n_{ik} the total corresponding read count. The observed *expressed* proportion of the reference allele is
 479 then given by $\Upsilon_{ik} = \frac{y_{ik}}{n_{ik}}$.

480 To test for ASE we used the hierarchical Beta-Binomial model proposed by ³⁷, with
 481 modifications. We first modelled y_{ik} according to a Binomial distribution, $y_{ik} \sim \text{Binomial}(n_{ik}, \theta_{ik})$.
 482 Next we modelled the *a priori* distribution of the parameter θ_{ik} with a Beta distribution,
 483 $\theta_{ik} \sim \text{Beta}(\alpha_{ik}, \beta_{ik})$, where α_{ik} and β_{ik} indicate the genomic dose of the reference and alternative
 484 alleles, respectively. Because these doses are unknown, we obtained estimates by approximating the
 485 observed genomic allele proportions, considering a ploidy of 12. In that case, we set $\alpha_{ik} = [12 \times \Gamma_{ik}]$,
 486 where the brackets represent the integer closest to $12 \times \Gamma_{ik}$; we then have that $\beta_{ik} = 12 - \alpha_{ik}$.

487 According to this model, the posterior distribution of θ_{ik} is then $\text{Beta}(y_{ik} + \alpha_{ik}, n_{ik} - y_{ik} +$
 488 $\beta_{ik})$, but in practice we used $\text{Beta}(y_{ik} + \alpha_{ik} + 0.5, n_{ik} - y_{ik} + \beta_{ik} + 0.5)$ to avoid zero counts. For
 489 each site we obtained the highest density interval (HDI) for this posterior distribution with the R
 490 package HDINTERVAL V0.2.2 ⁷⁸. To account for the large number of genotype $\times$ SNV combinations, we
 491 applied the Bonferroni correction ⁷⁹ to the interval mass, with a global significance level of 0.05 and
 492 considering an approximation of 100,000 independent tests. This is because neighbouring SNVs in a
 493 given gene are not independent, such that the number of tests must be between the number of
 494 genes and the number of positions tested.

495 Finally, to call an SNV as showing significant ASE we checked whether the HDI included the
 496 observed genomic proportion Γ of the reference allele. Because this proportion is an estimate and
 497 thus subject to random variation, instead of using a point estimate we used the 95% confidence
 498 interval for the binomial distribution calculated with the Wilson method ⁸⁰. If there was no overlap

499 between this confidence interval and the posterior HDI, we deemed the corresponding SNV to show
 500 significant allele-specific expression.

501 *Functional enrichment tests*

502 To search for enrichment of gene groups among those with ASE we first used ORTHOFINDER
 503 v.2.3.12^{81,82} to find clusters of genes conserved in the Liliopsida. We downloaded the proteomes of
 504 *Aegilops tauschii* v4.0, *Ananas comosus* v3, *Brachypodium distachyon* v3.1, *Dioscorea rotundata* v1.0,
 505 *Eragrostis curvula* v1.0, *Hordeum vulgare* v2, *Leersia perrieri* v1.4, *Musa acuminata* v1, *Miscanthus*
 506 *sinensis* v7.1, *Oropetium thomaeum* v1.0, *Oryza sativa* v7.0, *Panicum virgatum* v5.1, *Saccharum*
 507 *spontaneum*, *Setaria italica* v2.2, *Sorghum bicolor* v3.1.1, *Triticum aestivum* v2.2 and *Zea mays* v4
 508 from Phytozome v13⁶⁵ and Ensembl Plants 49⁶⁶. We ran Orthofinder with default parameters and
 509 identified clusters of genes with at least one representative in each of these species. Based on this
 510 analysis we also identified groups of paralogs that are exclusive to sorghum, the genome we used as
 511 reference, but absent from all other species. In parallel, we downloaded ORTHODB v10.1⁸³ to find
 512 genes that are present in single copy in sorghum, rice and Brachypodium.

513 For enrichment tests we filtered for SNVs with at least 50 genomic reads and 10 or more
 514 RNA-seq reads, and considered that genes showed ASE when they included at least two SNVs with
 515 significant ASE. For each genotype and internode combination we tested for enrichment using
 516 Fisher's exact test, and obtained FDR-corrected *p*-values according to⁸⁴. In all cases we tested for
 517 enrichment among Liliopsida-conserved, single copy and sorghum-exclusive paralogous genes.

518 We also searched for enriched Gene Ontology terms among genes with ASE^{47,48}. First, to
 519 each SNV present within an annotated gene we assigned the GO terms associated with that gene.
 520 Again we filtered to keep SNVs with a minimum of 50 genomic reads and 10 RNA-seq reads, and
 521 also removed genes with less than 10 SNVs. For each gene we calculated the median absolute

522 deviation between the expressed allele ratio and the genomic allele ratio of its corresponding
523 filtered SNVs. We then ranked the genes according to this median deviation and used
524 GSEAPRERANKED V4.0.3^{45,46} to test for enrichment of functional terms. We tested GO terms with five
525 or more genes, with 10,000 permutations and no collapsing of gene identifiers.

526 Acknowledgements

527 We wish to acknowledge The University of Queensland's Research Computing Centre (RCC)
528 for its support in this research. We also acknowledge the Center for Functional Genomics – ESALQ
529 for its computational infrastructure. This study was financed in part by the Coordenação de
530 Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – grants CAPES-PRINT
531 88887.466432/2019-00 to GRAM and CAPES-PRINT 88887.367965/2019-00 to FHC. Australian
532 Research Council and Sugar Research Australia provided funding to RJH and FCB.

533 Data Availability

534 Whole genome sequencing reads for the sugarcane hybrids have been deposited in the
535 National Center for Biotechnology Information (NCBI) under BioProject number PRJNA733812. The
536 RNA-seq reads used in this study are available in the European Nucleotide Archive (ENA) at EMBL-
537 EBI under accession number PRJEB444480.

538

539 **References**

- 540 1. Henry, R. J. Evaluation of plant biomass resources available for replacement of fossil oil. *Plant*
541 *Biotechnol. J.* **8**, 288–293 (2010).
- 542 2. OECD/FAO. Sugar. in *OECD-FAO Agricultural Outlook 2019-2028* 154–165 (OECD Publishing,
543 Paris/Food and Agriculture Organization of the United Nations, Rome, 2019).
544 doi:10.4060/CA4076EN
- 545 3. Souza, G. M. *et al.* The role of bioenergy in a climate-changing world. *Environ. Dev.* **23**, 57–64
546 (2017).
- 547 4. Goldemberg, J. Ethanol for a sustainable energy future. *Science (80-.)*. **315**, 808–810 (2007).
- 548 5. Wang, J., Nayak, S., Koch, K. & Ming, R. Carbon partitioning in sugarcane (*Saccharum* species).
549 *Front. Plant Sci.* **4**, 201 (2013).
- 550 6. D'Hont, A., Ison, D., Alix, K., Roux, C. & Glaszmann, J. C. Determination of basic chromosome
551 numbers in the genus *Saccharum* by physical mapping of ribosomal RNA genes. *Genome* **41**,
552 221–225 (1998).
- 553 7. Kim, C. *et al.* Comparative analysis of *Miscanthus* and *Saccharum* reveals a shared whole-
554 genome duplication but different evolutionary fates. *Plant Cell* **26**, 2420–2429 (2014).
- 555 8. D'Hont, A. *et al.* Characterisation of the double genome structure of modern sugarcane
556 cultivars (*Saccharum* spp.) by molecular cytogenetics. *Mol. Gen. Genet.* **250**, 405–413 (1996).
- 557 9. Grivet, L. & Arruda, P. Sugarcane genomics: depicting the complex genome of an important
558 tropical crop. *Curr. Opin. Plant Biol.* **5**, 122–127 (2002).
- 559 10. Piperidis, G., Piperidis, N. & D'Hont, A. Molecular cytogenetic investigation of chromosome
560 composition and transmission in sugarcane. *Mol. Genet. Genomics* **284**, 65–73 (2010).

- 561 11. D'Hont, A. & Glaszmann, J.-C. Sugarcane genome analysis with molecular markers, a first
562 decade of research. *Proc. Int. Soc. Sugar Cane Technol.* **24**, 556–559 (2001).
- 563 12. Le Cunff, L. *et al.* Diploid/polyploid syntenic shuttle mapping and haplotype-specific
564 chromosome walking toward a rust resistance gene (Bru1) in highly polyploid sugarcane (2n
565 ~ 12x ~ 115). *Genetics* **180**, 649–660 (2008).
- 566 13. Piperidis, N. & D'Hont, A. Sugarcane genome architecture decrypted with chromosome-
567 specific oligo probes. *Plant J.* **103**, 2039–2051 (2020).
- 568 14. Pompidor, N. *et al.* Three founding ancestral genomes involved in the origin of sugarcane.
569 *Ann. Bot.* **XX**, 1–14 (2021).
- 570 15. Yadav, S. *et al.* Accelerating genetic gain in sugarcane breeding using genomic selection.
571 *Agronomy* **10**, 1–21 (2020).
- 572 16. Kandel, R., Yang, X., Song, J. & Wang, J. Potentials, challenges, and genetic and genomic
573 resources for sugarcane biomass improvement. *Front. Plant Sci.* **9**, 151 (2018).
- 574 17. Diniz, A. L. *et al.* Genomic resources for energy cane breeding in the post genomics era.
575 *Comput. Struct. Biotechnol. J.* **17**, 1404–1414 (2019).
- 576 18. Thirugnanasambandam, P. P., Hoang, N. V & Henry, R. J. The Challenge of Analyzing the
577 Sugarcane Genome. *Front. Plant Sci.* **9**, 616 (2018).
- 578 19. Garsmeur, O. *et al.* A mosaic monoploid reference sequence for the highly complex genome
579 of sugarcane. *Nat. Commun.* **9**, 2638 (2018).
- 580 20. Zhang, J. *et al.* Allele-defined genome of the autopolyploid sugarcane *Saccharum*
581 *spontaneum* L. *Nat. Genet.* **50**, 1565–1573 (2018).
- 582 21. Souza, G. M. *et al.* Assembly of the 373k gene space of the polyploid sugarcane genome

- reveals reservoirs of functional diversity in the world's leading biomass crop. *Gigascience* **8**, 1–18 (2019).
22. Dillon, S. L. *et al.* Domestication to crop improvement: Genetic resources for Sorghum and Saccharum (Andropogoneae). *Ann. Bot.* **100**, 975–989 (2007).
 23. Wang, J. *et al.* Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes. *BMC genomics* **11**, 261 (2010).
 24. Comai, L. Genetic and epigenetic interactions in allopolyploid plants. *Plant Mol. Biol.* **43**, 387–399 (2000).
 25. Renny-Byfield, S. & Wendel, J. F. Doubling down on genomes: Polyploidy and crop plants. *Am. J. Bot.* **101**, 1711–1725 (2014).
 26. Yant, L. & Bomblies, K. Genome management and mismanagement—cell-level opportunities and challenges of whole-genome duplication. *Genes Dev.* **29**, 2405–2419 (2015).
 27. Feldman, M. & Levy, A. A. Genome evolution in allopolyploid wheat—a revolutionary reprogramming followed by gradual changes. *J. Genet. Genomics* **36**, 511–518 (2009).
 28. Baduel, P., Bray, S., Vallejo-Marin, M., Kolář, F. & Yant, L. The 'Polyploid Hop': Shifting challenges and opportunities over the evolutionary lifespan of genome duplications. *Front. Ecol. Evol.* **6**, 117 (2018).
 29. Chen, Z. J. & Ni, Z. Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. *BioEssays* **28**, 240–252 (2006).
 30. Gaur, U., Li, K., Mei, S. & Liu, G. Research progress in allele-specific expression and its regulatory mechanisms. *J. Appl. Genet.* **54**, 271–283 (2013).
 31. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best

- 605 practices for data processing in allelic expression analysis. *Genome Biol.* **16**, 195 (2015).
- 606 32. Chen, C. *et al.* Characterization of imprinted genes in rice reveals conservation of regulation
607 and imprinting with other plant species. *Plant Physiol.* **177**, 1754–1771 (2018).
- 608 33. Zhang, X. & Borevitz, J. O. Global analysis of allele-specific expression in *Arabidopsis thaliana*.
609 *Genetics* **182**, 943–954 (2009).
- 610 34. Springer, N. M. & Stupar, R. M. Allele-specific expression patterns reveal biases and embryo-
611 specific parent-of-origin effects in hybrid maize. *Plant Cell* **19**, 2391–2402 (2007).
- 612 35. Pham, G. M. *et al.* Extensive genome heterogeneity leads to preferential allele expression and
613 copy number-dependent expression in cultivated potato. *Plant J.* **92**, 624–637 (2017).
- 614 36. Powell, J. J. *et al.* The defence-associated transcriptome of hexaploid wheat displays
615 homoeolog expression and induction bias. *Plant Biotechnol. J.* **15**, 533–543 (2017).
- 616 37. Correr, F. H., Furtado, A., Garcia, A. A. F., Henry, R. J. & Margarido, G. R. A. A hierarchical
617 Bayesian model to assess allele-specific expression in mixed-ploidy species reveals expression
618 biases in sugarcane. *Plant J.* (submitted) (2021).
- 619 38. Sforça, D. A. *et al.* Gene duplication in the sugarcane genome: A case study of allele
620 interactions and evolutionary patterns in two genic regions. *Front. Plant Sci.* **10**, 553 (2019).
- 621 39. Vilela, M. de M. *et al.* Analysis of Three Sugarcane Homo/Homeologous Regions Suggests
622 Independent Polyploidization Events of *Saccharum officinarum* and *Saccharum spontaneum*.
623 *Genome Biol. Evol.* **9**, 266–278 (2017).
- 624 40. Cai, M. *et al.* Allele specific expression of Dof genes responding to hormones and abiotic
625 stresses in sugarcane. *PLoS One* **15**, e0227716 (2020).
- 626 41. Bell, G. D. M., Kane, N. C., Rieseberg, L. H. & Adams, K. L. RNA-seq analysis of allele-specific

- 627 expression, hybrid effects, and regulatory divergence in hybrids compared with their parents
628 from natural populations. *Genome Biol. Evol.* **5**, 1309–1323 (2013).
- 629 42. Whittaker, A. & Botha, F. C. Carbon partitioning during sucrose accumulation in sugarcane
630 internodal tissue. *Plant Physiol.* **115**, 1651–1659 (1997).
- 631 43. Botha, F. C. & Black, K. G. Sucrose phosphate synthase and sucrose synthase activity during
632 maturation of internodal tissue in sugarcane. *Aust. J. Plant Physiol.* **27**, 81–85 (2000).
- 633 44. Perlo, V., Botha, F. C., Furtado, A., Hodgson-Kratky, K. & Henry, R. J. Metabolic changes in the
634 developing sugarcane culm associated with high yield and early high sugar content. *Plant*
635 *Direct* **4**, e00276 (2020).
- 636 45. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for
637 interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550
638 (2005).
- 639 46. Mootha, V. K. *et al.* PGC-1 α -responsive genes involved in oxidative phosphorylation are
640 coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
- 641 47. The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOLD mine. *Nucleic*
642 *Acids Res.* **49**, D325–D334 (2021).
- 643 48. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29
644 (2000).
- 645 49. Held, B. M., Wang, H., John, I., Wurtele, E. S. & Colbert, J. T. An mRNA putatively coding for an
646 O-methyltransferase accumulates preferentially in maize roots and is located predominantly
647 in the region of the endodermis. *Plant Physiol.* **102**, 1001–1008 (1993).
- 648 50. Bosch, M., Mayer, C. D., Cookson, A. & Donnison, I. S. Identification of genes involved in cell

- 649 wall biogenesis in grasses by differential gene expression profiling of elongating and non-
 650 elongating maize internodes. *J. Exp. Bot.* **62**, 3545–3561 (2011).
- 651 51. Scott-Craig, J. S., Casida, J. E., Poduje, L. & Walton, J. D. Herbicide Safener-Binding Protein of
 652 Maize: Purification, Cloning, and Expression of an Encoding cDNA. *Plant Physiol.* **116**, 1083–
 653 1089 (1998).
- 654 52. Rivas, M. A. *et al.* Effect of predicted protein-truncating genetic variants on the human
 655 transcriptome. *Science (80-.)*. **348**, 666–669 (2015).
- 656 53. Margarido, G. R. A. & Heckerman, D. ConPADE: genome assembly ploidy estimation from
 657 next-generation sequencing data. *PLoS Comput. Biol.* **11**, e1004229 (2015).
- 658 54. Gerard, D., Ferrão, L. F. V., Garcia, A. A. F. & Stephens, M. Genotyping Polyploids from Messy
 659 Sequencing Data. *Genetics* **210**, 789–807 (2018).
- 660 55. Geijn, B. van de, McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for
 661 robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).
- 662 56. Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J. & Akey, J. M. A powerful and flexible
 663 statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq
 664 data. *Genome Res.* **21**, 1728–1737 (2011).
- 665 57. Scheben, A., Batley, J. & Edwards, D. Genotyping-by-sequencing approaches to characterize
 666 crop genomes: choosing the right tool for the right application. *Plant Biotechnol. J.* **15**, 149–
 667 161 (2017).
- 668 58. Tameling, W. I. L. *et al.* Mutations in the NB-ARC domain of I-2 that impair ATP hydrolysis
 669 cause autoactivation. *Plant Physiol.* **140**, 1233–1245 (2006).
- 670 59. DeYoung, B. J. & Innes, R. W. Plant NBS-LRR proteins in pathogen sensing and host defense.

- 671 *Nat. Immunol.* **7**, 1243–1249 (2006).
- 672 60. Cheavegatti-Gianotto, A. *et al.* Sugarcane (*Saccharum X officinarum*): A Reference Study for
673 the Regulation of Genetically Modified Cultivars in Brazil. *Trop. Plant Biol.* **4**, 62–89 (2011).
- 674 61. Rott, P. C., Girard, J.-C. & Comstock, J. C. Impact of pathogen genetics on breeding for
675 resistance to sugarcane diseases. *Proc. XXVIII ISSCT (International Soc. Sugar Cane Technol.*
676 *Congr.* **28**, (2013).
- 677 62. Furtado, A. DNA Extraction from Vegetative Tissue for Next-Generation Sequencing. in *Cereal*
678 *Genomics: Methods and Protocols* (eds. Henry, R. J. & Furtado, A.) 1–5 (Humana Press, 2014).
679 doi:10.1007/978-1-62703-715-0_1
- 680 63. Paterson, A. H. *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature*
681 **457**, 551–556 (2009).
- 682 64. McCormick, R. F. *et al.* The *Sorghum bicolor* reference genome: improved assembly, gene
683 annotations, a transcriptome atlas, and signatures of genome organization. *Plant J.* **93**, 338–
684 354 (2018).
- 685 65. Goodstein, D. M. *et al.* Phytozome: A comparative platform for green plant genomics. *Nucleic*
686 *Acids Res.* **40**, D1178–D1186 (2012).
- 687 66. Yates, A. D. *et al.* Ensembl 2020. *Nucleic Acids Res.* **48**, D682–D688 (2020).
- 688 67. Durinck, S. *et al.* BioMart and Bioconductor: A powerful link between biological databases and
689 microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).
- 690 68. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of
691 genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191
692 (2009).

- 693 69. R Core Team. R: A Language and Environment for Statistical Computing. (2020). Available at:
694 <https://www.r-project.org/>.
- 695 70. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**,
696 357–9 (2012).
- 697 71. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079
698 (2009).
- 699 72. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-
700 generation DNA sequencing data. *Nat. Genet.* **43**, 491–8 (2011).
- 701 73. Obenchain, V. *et al.* VariantAnnotation: A Bioconductor package for exploration and
702 annotation of genetic variants. *Bioinformatics* **30**, 2076–2078 (2014).
- 703 74. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide
704 polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2;
705 iso-3. *Fly (Austin)*. **6**, 80–92 (2012).
- 706 75. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence
707 data. *Bioinformatics* **30**, 2114–2120 (2014).
- 708 76. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing
709 and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
- 710 77. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: A fast spliced aligner with low memory
711 requirements. *Nat. Methods* **12**, 357–360 (2015).
- 712 78. Meredith, M. & Kruschke, J. HDInterval: Highest (Posterior) Density Intervals. (2020).
- 713 79. Bonferroni, C. E. Teoria statistica delle classi e calcolo delle probabilità. *Pubbl. del R Ist. Super.*
714 *di Sci. Econ. e Commer. di Firenze* **8**, 3–62 (1936).

- 715 80. Wilson, E. B. Probable Inference, the Law of Succession, and Statistical Inference. *J. Am. Stat.*
716 *Assoc.* **22**, 209–212 (1927).
- 717 81. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome
718 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157
719 (2015).
- 720 82. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative
721 genomics. *Genome Biol.* **20**, 238 (2019).
- 722 83. Kriventseva, E. V *et al.* OrthoDB v10: sampling the diversity of animal, plant, fungal, protist,
723 bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic*
724 *Acids Res.* **47**, D807–D811 (2019).
- 725 84. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful
726 approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).

727

Table 1. Results of the allele-specific expression (ASE) test. For each combination of sugarcane genotype and internode we show the total number of informative sites, which are heterozygous with at least 50 genomic reads and 10 RNA-seq reads, and those with significant ASE. Sampled genes are those with two or more informative variant sites.

Genotype	Internode	Sites tested	Sites with ASE	Sampled genes	Genes with ASE
KQ228	C1 In5	784,678	50,969	19,254	5,652 (29.35%)
	C1 In8	776,315	56,956	18,544	5,881 (31.71%)
	C2 In5	670,019	38,142	17,644	4,161 (23.58%)
	C2 In8	779,532	54,029	18,370	5,577 (30.36%)
	C2 InEx-5	788,762	54,494	18,151	5,660 (31.18%)
KQB09	C1 In5	635,981	31,955	18,429	3,799 (20.61%)
	C1 In8	682,963	43,452	17,918	4,637 (25.88%)
	C2 In5	740,234	34,953	18,698	4,060 (21.71%)
	C2 In8	708,982	42,213	17,726	4,476 (25.25%)
	C2 InEx-5	769,011	47,203	17,393	4,873 (28.02%)
MQ239	C1 In5	591,330	32,415	18,375	3,789 (20.62%)
	C1 In8	640,136	36,967	17,448	4,025 (23.07%)
	C2 In5	673,876	38,622	17,813	4,059 (22.79%)
	C2 In8	708,504	42,413	17,807	4,462 (25.06%)
	C2 InEx-5	644,549	30,013	17,125	3,377 (19.72%)
Q155	C1 In5	599,518	28,372	18,052	3,402 (18.85%)
	C1 In8	670,021	40,431	17,776	4,391 (24.70%)
	C2 In5	695,515	38,248	18,057	4,167 (23.08%)
	C2 In8	708,493	36,893	17,750	4,074 (22.95%)
	C2 InEx-5	714,418	37,491	17,522	4,054 (23.14%)
Q186	C1 In5	684,824	37,348	18,729	4,224 (22.55%)
	C1 In8	659,836	35,153	17,478	3,833 (21.93%)
	C2 In5	577,941	27,231	16,653	2,971 (17.84%)
	C2 In8	746,505	41,164	17,836	4,342 (24.34%)
	C2 InEx-5	599,414	25,547	16,538	2,849 (17.23%)
SRA1	C1 In5	512,827	19,154	16,936	2,311 (13.65%)
	C1 In8	648,316	37,685	17,438	3,976 (22.80%)
	C2 In5	758,633	41,910	18,393	4,482 (24.37%)
	C2 In8	741,704	43,821	18,188	4,602 (25.30%)
	C2 InEx-5	655,869	29,526	17,204	3,215 (18.69%)
SRA5	C1 In5	628,813	33,587	18,109	3,967 (21.91%)
	C1 In8	788,206	52,306	18,292	5,467 (29.89%)
	C2 In5	851,295	55,743	18,879	5,902 (31.26%)
	C2 In8	815,312	52,494	18,556	5,642 (30.41%)
	C2 InEx-5	798,053	49,362	17,898	5,264 (29.41%)

Fig. 1. Relationship between expressed and genomic proportions of the reference allele for genotype KQ228. Different internodes are shown in separate panels. The red line indicates the null hypothesis of perfect identity between the expressed proportion of the reference allele (Υ , in the Y-axis) and the corresponding genomic proportion (denoted by Γ , X-axis). Each point represents one single nucleotide variant, and lighter colours indicate a higher density of points. The smoothed trend of observed points is shown in pink. Notice the majority of sites with high proportion of the reference allele in the genome. To highlight the discrete nature of clusters we only show high-depth sites, with 1,000 or more genomic reads and 80 or more RNA-seq reads.

Fig. 2. 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.

Fig. 3. Percentage of genes with allele-specific expression detected in common for all genotype × internode combinations. We found higher similarity among different internodes of the same genotype (hot colours in the diagonal blocks), and there was also similarity between the same internode of different genotypes (the light orange cells off the main diagonal).

Fig. 4. Comparison between KQ228 and SRA5, the two genotypes sequenced at higher depth. A) Relationship between genomic allele ratios (doses) for shared variants, showing overall agreement between the two genotypes. B) Relationship between allele proportions in their expressed

759 sequences. C) Comparison of expression bias for shared SNVs with significant ASE, showing
760 prevalence of higher expression of the reference allele in both genotypes.

Fig. 5. Enrichment results of genes with allele-specific expression. A) Single nucleotide variants (SNVs) in genes annotated with ontology term *defence response* (GO:0006952). B) SNVs in genes annotated with ontology term *UDP-glycosyltransferase activity* (GO:0008194). C) SNVs predicted to have high functional impact. In all cases, note the excess of SNVs with exclusive expression of the reference allele (horizontal line at $Y = 1$). D) Underrepresentation of single copy genes among those with ASE (odds ratio < 1); overrepresentation of core genes, conserved in Liliopsida (odds ratio > 1). Multiple points for the same genotype indicate different internodes.

770 **Supplementary Table 1.** Whole-genome sequencing and read trimming statistics. Alignment rates
 771 shown are against the sorghum genome.

Genotype	Raw reads (millions)	Raw bases (billions)	High quality reads (millions)	High quality bases (billions)	Alignment rate
KQ228	7,984.80	1,205.70	7,736.15 (96.89 %)	1,133.90 (94.04 %)	37.66 %
SRA5	7,906.10	1,193.82	7,651.47 (96.78 %)	1,121.77 (93.97 %)	38.48 %
Q155	1,595.45	240.91	1,525.27 (95.60 %)	223.38 (92.72 %)	37.31 %
KQB09	1,616.19	244.04	1,541.7 (95.39 %)	225.91 (92.57 %)	37.84 %
SRA1	1,618.08	244.33	1,551.67 (95.90 %)	227.25 (93.01 %)	37.24 %
MQ239	1,523.06	229.98	1,447.08 (95.01 %)	211.52 (91.97 %)	37.10 %
Q186	1,608.74	242.92	1,528.13 (94.99 %)	222.85 (91.74 %)	37.15 %

Supplementary Table 2. RNA-seq preprocessing and alignment statistics. Alignment rates against the sorghum genome.

Genotype	Collection	Internode	Raw reads	High quality reads	Alignment rate
KQ228	C1	In5	63,516,146	53,340,070 (83.98 %)	83.50 %
			79,526,358	64,930,414 (81.65 %)	82.11 %
			76,554,790	62,904,286 (82.17 %)	84.39 %
		In8	76,152,860	63,111,436 (82.87 %)	83.43 %
			73,443,866	61,165,200 (83.28 %)	83.87 %
			65,575,500	54,032,056 (82.40 %)	84.29 %
	C2	In5	63,893,630	32,312,360 (50.57 %)	84.24 %
			73,479,056	56,346,352 (76.68 %)	85.39 %
			75,032,818	57,091,702 (76.09 %)	83.46 %
		In8	77,725,398	63,551,444 (81.76 %)	80.68 %
			63,070,770	49,401,954 (78.33 %)	80.02 %
			58,260,760	48,306,654 (82.91 %)	82.84 %
	InEx-5		67,083,624	52,099,898 (77.66 %)	73.63 %
			74,335,192	62,727,224 (84.38 %)	74.20 %
			71,197,488	59,525,388 (83.61 %)	77.14 %
SRA5	C1	In5	63,399,486	52,771,562 (83.24 %)	84.54 %
			56,016,590	46,294,142 (82.64 %)	83.39 %
			62,370,014	51,027,974 (81.81 %)	84.87 %
		In8	61,376,352	50,873,230 (82.89 %)	82.77 %
			64,792,956	53,279,652 (82.23 %)	82.80 %
			74,020,306	60,540,584 (81.79 %)	82.57 %
	C2	In5	76,749,354	64,855,364 (84.50 %)	85.91 %
			79,188,660	65,830,172 (83.13 %)	83.93 %
			60,435,148	48,432,216 (80.14 %)	81.26 %
		In8	77,105,510	64,382,752 (83.50 %)	81.16 %
			66,308,550	55,367,412 (83.50 %)	82.86 %
			72,417,228	60,088,340 (82.98 %)	83.58 %
	InEx-5		67,287,950	50,794,552 (75.49 %)	76.00 %
			69,563,210	53,992,048 (77.62 %)	76.60 %
			74,831,050	63,088,510 (84.31 %)	79.66 %
Q155	C1	In5	70,541,218	58,964,774 (83.59 %)	82.27 %
			63,114,730	50,472,396 (79.97 %)	78.29 %
			66,153,896	55,641,124 (84.11 %)	84.63 %
		In8	58,515,114	47,072,744 (80.45 %)	80.83 %
			75,307,846	61,748,452 (81.99 %)	82.74 %
			82,714,180	67,775,484 (81.94 %)	84.56 %
	C2	In5	70,404,020	58,131,740 (82.57 %)	84.02 %
			61,889,754	48,432,258 (78.26 %)	83.59 %
			74,505,566	57,899,820 (77.71 %)	85.05 %
		In8	71,417,736	59,451,024 (83.24 %)	80.90 %

KQB09			69,679,156	54,216,066 (77.81 %)	81.42 %
			68,475,280	52,720,622 (76.99 %)	76.64 %
		InEx-5	69,258,394	57,344,844 (82.80 %)	77.53 %
			70,703,414	57,602,860 (81.47 %)	79.26 %
			65,646,298	49,946,400 (76.08 %)	73.34 %
	C1	In5	68,393,442	54,669,414 (79.93 %)	84.64 %
			71,101,340	58,461,218 (82.22 %)	81.80 %
			74,372,296	61,970,994 (83.33 %)	84.45 %
		In8	70,208,596	58,271,460 (83.00 %)	82.85 %
			79,095,960	63,816,580 (80.68 %)	82.86 %
			77,780,894	63,144,840 (81.18 %)	84.80 %
	C2	In5	85,113,084	70,321,428 (82.62 %)	81.40 %
			73,688,532	61,365,220 (83.28 %)	84.83 %
			63,771,802	50,692,570 (79.49 %)	82.32 %
		In8	72,910,004	60,577,168 (83.08 %)	83.68 %
			64,355,436	52,636,932 (81.79 %)	83.83 %
			67,782,386	52,451,172 (77.38 %)	83.84 %
SRA1		In5	78,754,954	65,818,624 (83.57 %)	80.59 %
			83,012,470	62,496,796 (75.29 %)	82.89 %
			66,853,872	52,560,458 (78.62 %)	82.16 %
		InEx-5	61,005,106	50,504,258 (82.79 %)	83.36 %
			62,023,168	50,131,538 (80.83 %)	83.25 %
			62,654,368	50,209,682 (80.14 %)	83.45 %
	C1	In5	59,545,458	50,144,122 (84.21 %)	83.34 %
			65,250,582	53,341,278 (81.75 %)	82.73 %
			67,936,708	54,952,916 (80.89 %)	83.84 %
		In8	73,407,066	60,703,188 (82.69 %)	83.00 %
			74,342,666	61,774,744 (83.09 %)	82.31 %
			63,529,892	50,145,954 (78.93 %)	85.22 %
	C2	In5	79,961,544	65,018,698 (81.31 %)	83.36 %
			65,051,192	50,962,402 (78.34 %)	82.27 %
			72,615,990	60,883,324 (83.84 %)	82.84 %
		InEx-5	74,938,264	61,684,104 (82.31 %)	75.87 %
			62,497,930	43,001,060 (68.80 %)	76.47 %
			58,096,448	40,901,050 (70.40 %)	77.56 %
MQ239		In5	68,536,344	57,160,794 (83.40 %)	83.82 %
			74,847,252	62,599,658 (83.64 %)	84.59 %
			75,934,864	64,118,832 (84.44 %)	84.30 %
		In8	74,913,848	61,791,260 (82.48 %)	83.05 %
			65,440,494	52,293,002 (79.91 %)	82.26 %
			62,502,938	50,976,270 (81.56 %)	79.28 %
	C1	In5	59,410,854	46,745,682 (78.68 %)	83.98 %
			64,669,344	50,797,130 (78.55 %)	85.98 %
			69,532,422	58,545,028 (84.20 %)	85.43 %
		In8	70,905,250	60,033,102 (84.67 %)	83.64 %
			62,224,882	47,999,144 (77.14 %)	81.81 %
			71,626,762	56,685,896 (79.14 %)	83.57 %

Q186	InEx-5	71,848,986	58,212,228 (81.02 %)	75.64 %
		59,872,388	49,040,456 (81.91 %)	72.41 %
		60,431,746	46,400,672 (76.78 %)	78.40 %
	In5	74,543,176	61,870,078 (83.00 %)	83.57 %
		78,870,528	64,737,830 (82.08 %)	81.87 %
		78,035,396	65,065,928 (83.38 %)	84.59 %
	In8	75,366,534	62,723,942 (83.23 %)	83.22 %
		63,987,526	50,696,984 (79.23 %)	82.71 %
		65,154,614	53,702,984 (82.42 %)	82.48 %
	In5	66,536,168	55,550,046 (83.49 %)	84.67 %
		64,176,926	46,509,140 (72.47 %)	82.70 %
		58,481,330	33,879,064 (57.93 %)	83.66 %
	In8	78,270,556	62,932,100 (80.40 %)	81.48 %
		62,664,308	48,560,828 (77.49 %)	82.20 %
		68,572,542	54,759,398 (79.86 %)	83.49 %
	InEx-5	64,309,392	37,186,144 (57.82 %)	79.36 %
		67,288,592	52,024,078 (77.31 %)	77.07 %
		76,014,900	58,374,006 (76.79 %)	76.26 %

777 **Supplementary Table 3.** Genes showing allele-specific expression for each genotype and internode.
778 Genes are in the rows and treatments in the columns. Each cell of the incidence matrix contains 1 if
779 the gene showed significant ASE for a given treatment; 0 if there was no evidence of ASE; NA if there
780 were no informative variants.

781

782 See attached file with Supplementary Table 3.

783 **Supplementary Table 4.** Functional enrichment of genes with allele-specific expression according
784 to Gene Ontology terms. Functional terms are in the rows and treatments are in the columns.
785 Significant p -values are shown ($p < 0.05$) after false discovery rate correction for multiple tests.

786

787 See attached file with Supplementary Table 4.

Supplementary Fig. 1. Distribution of the median depth of coverage when aligning genomic reads from genotype KQ228 to the sorghum genome. A) Distribution for different gene features: CDS (coding sequence), exons and the entire gene (including introns). Note the peak at roughly 1200 X, which is 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.

Supplementary Fig. 2. 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.

Supplementary Fig. 3. 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).

Supplementary Fig. 4. 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 Supplementary Table 4.

Supplementary Fig. 5. 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. C) Distribution of the deviation between RNA and DNA allele ratios (expression bias). Positive values indicate higher expression of the reference allele.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryTable3.csv](#)
- [SupplementaryTable4.csv](#)