

Major impacts of widespread structural variation on sorghum

Zhihai Zhang¹, Joao Paulo Gomes Viana², Bosen Zhang², Kimberly K.O. Walden³, Hans Müller Paul¹, Stephen P Moose^{1,2}, Geoff Morris⁴, Chris Daum⁵, Kerrie W Barry⁵, Nadia Shakoor⁶, Matthew E Hudson^{1,2*}.

¹ DOE Center for Advanced Bioenergy and Bioproducts Innovation (CABBI), University of Illinois at Urbana-Champaign, IL 61801, USA. ² Department of Crop Sciences, University of Illinois at Urbana-Champaign, 1102S Goodwin Ave, Urbana, IL 61801, USA. ³ High Performance Computing in Biology, Carver Biotechnology Center, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. ⁴ Department of Soil & Crop Science, Colorado State University, Plant Science Building, Fort Collins, CO, 80523, USA. ⁵ United States Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA. ⁶ Donald Danforth Plant Science Center, St. Louis, MO 63132, USA.

* Corresponding author, email: mhudson@illinois.edu

Abstract

Genetic diversity is critical to crop breeding and improvement, and dissection of the genomic variation underlying agronomic traits can both assist breeding and give insight into basic biological mechanisms. While recent genome analyses in plants reveal many structural variants (SVs), most current studies of crop genetic variation are dominated by single nucleotide polymorphisms (SNPs). The extent of the impact of SVs on global trait variation, and their utility in genome-wide selection, is not yet understood. In this study, we built an SV dataset based on whole-genome resequencing of diverse sorghum lines (n = 363), validated the correlation of photoperiod sensitivity and variety type, and identified SV hotspots underlying the divergent evolution of cellulosic and sweet sorghum. In addition, we demonstrated the complementary contribution of SVs for heritability of traits related to sorghum adaptation. Importantly, inclusion of SV polymorphisms in association studies revealed genotype-phenotype associations not observed with SNPs alone. Three-way genome-wide association studies (GWAS) based on whole-genome SNP, SV, and integrated SNP+SV datasets demonstrated substantial associations between SVs and sorghum traits. Addition of SVs to GWAS substantially increased heritability estimates for some traits, indicating their important contribution to functional allelic variation at

29 the genome level. Our discovery of the widespread impacts of SVs on heritable gene expression
30 variation could render a plausible mechanism for their disproportionate impact on phenotypic
31 variation. This study expands our knowledge of SVs and emphasizes the extensive impacts of
32 SVs on sorghum.

33

34 **Introduction**

35 High-throughput sequencing technologies have sped up the process of discovery for natural
36 genetic variation. However, as a consequence of limited read length and variant calling
37 algorithms, single nucleotide polymorphisms (SNPs) and small indels are disproportionately
38 overrepresented within characterized sequence variation (Audano et al. 2019). Nevertheless, a
39 growing number of research projects indicate that structural variations (SVs), including large
40 (>30bp) deletions, insertions, duplications, inversions, and translocations (Feuk et al. 2006),
41 greatly contribute to crop phenotypic diversity and selection for physiological and morphological
42 phenotypes (Alonge et al. 2020; Li et al. 2020). Two major SV classes have been proposed to
43 explain how structural variations are formed and how they impact phenotypes. The first involves
44 genome rearrangement, such as inversions and translocations; the second includes large deletions,
45 insertions, and duplications, collectively referred to as copy number variations (CNVs) (Alkan et
46 al. 2011; Scherer et al. 2007). Because structural variations are diverse and influence gene
47 sequence and expression via a myriad of mechanisms, it has been challenging to assess the
48 impact of SVs systematically and comprehensively. In addition, current sequencing and
49 detection technologies leave the bulk of SVs poorly resolved, so they are often not included in
50 studies of genome-wide variation.

51 Due to its low cost, mature and reliable technology, and proven high accuracy reads, second or
52 “next-generation” short read sequencing technologies are still the main technology for most
53 studies. Long-read sequencing techniques, such as PacBio HiFi and ultralong Oxford Nanopore
54 (ONT), are both more expensive and more demanding of DNA quantity and quality. Numerous
55 tools have been developed to detect SVs using paired-end short reads over the past decade. There
56 are primarily three strategies used in popular algorithms for SV calling based on short-read
57 sequencing: read-pair technologies, read-depth methods, and split-read approaches (Alkan et al.
58 2011). However, there is currently no individual algorithm that is able to successfully identify all
59 types of SVs across the entire range of sizes, as strategies display a diversity of strengths and
60 weaknesses in their ability to detect various types of SVs. Utilization of multiple algorithms
61 based on different strategies for SV detection has been proven a viable way to overcome this
62 issue. Zarate et al. found that reaching a consensus among multiple short-read SV callers can
63 lead to improved precision without significantly compromising sensitivity in human genome
64 (Zarate et al. 2020). Alonge et al. deployed three independent tools to call SVs from the short-
65 read alignments of 847 tomato accessions and successfully identified the diverse modern and
66 domesticated samples that maximize SV diversity (Alonge et al. 2020). In this study, we
67 developed an ensemble pipeline for SV calling based on five independent algorithms involving
68 different SV detection strategies: Sentieon (Kendig et al. 2019), which uses split-reads strategy
69 to call SVs and was also used for SNP calling in this study; DELLY (Rausch et al. 2012), which
70 uses paired-ends, split-reads and read-depth strategies to sensitively and accurately delineate SVs;
71 Smoove (<https://github.com/brentp/smoove>), which is an improved version of lumpy and
72 integrates the paired-end and split-read strategies; Manta (Chen et al. 2016), which combines

73 paired and split-read evidence during SV discovery; and CNVnator (Abyzov et al. 2011), which
74 utilizes read-depth methods.

75 The standard assumptions of genome-wide association studies (GWAS) include the concept that
76 each SNP used in the study will capture heritable variation via “tagging” any other SNPs, or SVs,
77 in the genome within the range of local linkage disequilibrium (LD) (Kruglyak 2008). For this
78 reason, it has been widely assumed that causative SVs will be detected in GWAS via being
79 “tagged” by adjacent SNPs in LD. Recent evidence has shed doubt on this assumption in plants,
80 due to the limited LD of many SVs with surrounding SNPs in soybean (Fliege et al. 2022) and
81 maize (Yang et al. 2019). For this reason, many of the effects of SVs on crop phenotypes may
82 still be unknown.

83 Sorghum (*Sorghum bicolor* (L.) Moench) is a versatile crop with wide adaptability and broad
84 applications. It has been selectively bred into different varieties for different end uses, such as
85 grain sorghum for human consumption, forage sorghum that is primarily for feeding livestock,
86 and sweet sorghum, which can be utilized as a food sweetener or for biofuel and chemical
87 production. These types have been created by selective breeding following sorghum
88 domestication in northern Africa about 10,000 years ago and its subsequent spread to a variety of
89 areas across Africa, India, the Middle East, and east Asia (Lobell et al. 2008; Morris et al. 2013a).
90 Each specific sorghum type is characterized by particular morphological and physiological
91 features. A better understanding of the genetic pathways and mechanisms that underpin these
92 features is essential for accelerating future sorghum breeding and improvement. Here, we aimed
93 to build an SV dataset based on whole-genome short-read resequencing of 363 sorghum lines
94 from the global Bioenergy Association Panel (BAP) (Brenton et al. 2016) using a fusion

95 workflow, investigate the impacts of SVs on sorghum genetics, and find new knowledge of
96 allelic variation that can be used in crop improvement.

97 **Results**

98 **Identification of Genome-wide Variations in the Bioenergy Association Panel**

99 In order to explore the genetics of SVs within sorghum germplasm, we utilized the Illumina
100 short-read whole-genome resequencing data from 363 global sorghum accessions in the BAP
101 (Brenton et al. 2016) (<https://terraref.org/>) (Supplemental Table S1). This panel was developed
102 and characterized as a set of racially, geographically, and phenotypically diverse lines aiming to
103 cover a significant portion of the genetic variation within sorghum (Hu et al. 2019). The panel
104 has been classified into three broad types: cellulosic, grain, and sweet (Brenton et al. 2016). The
105 mean sequencing depth is $\sim 29\times$ and the mean breadth of the coverage is $\sim 91\%$. Sorghum
106 BTx623 (v3.1.1) from Phytozome (<https://phytozome.jgi.doe.gov/>) was used as the reference
107 genome in SNP and SV calling. To enhance the accuracy and sensitivity of SV detection, five
108 inference software packages: Sentieon (v202010.01) (Kendig et al. 2019), DELLY (v0.8.1)
109 (Rausch et al. 2012), Smoove (<https://github.com/brentp/smoove>), manta (v1.6.0) (Chen et al.
110 2016) and CNVnator (v0.3.3) (Abyzov et al. 2011), involving different SV detection strategies,
111 were applied to the data. We conducted a simulation study to estimate recall and precision in SV
112 calling using various thresholds, considering SVs supported by one to five callers (Supplemental
113 Fig S1, Supplemental Results). Based on the simulation result, only SVs supported by at least
114 two callers were reported by our fusion workflow, and the two calls must agree on the type and
115 the strand of SV. A total of 7,162,000 filtered SNPs and 622,236 high-confidence SVs were
116 identified on 10 chromosomes, including 158,614 deletions (DEL), 18,028 duplications (DUP),

117 216 insertions (INS), 142,219 inversions (INV) and 303,159 translocations (TRA) (Supplemental
118 Fig S2A).

119 To validate the quality of the identified SVs, three new chromosome-scale *de novo* assemblies
120 (Supplemental Fig S2B-D, Supplemental Table S2) and two public whole genome sequence
121 assemblies available at Phytozome (<https://phytozome-next.jgi.doe.gov/>) for five BAP
122 accessions (PI 329545, PI 337680, PI 651495, Rio (Sorghum bicolor Rio v2.1) and RTx430
123 (Sorghum bicolor RTx430 v2.1)) were aligned to the standard reference genome (BTx623 v3.1.1)
124 and structural variants called by assembly comparison. SVs identified from whole-genome
125 alignment information were then compared with the SVs detected by the fusion workflow using
126 Illumina data. Overall, we observed a high percentage of overlapping fusion workflow calls with
127 assembly comparison for both DEL/INS and DUP and traceable breakpoints of TRA and INV.
128 We concluded that our fusion pipeline is sufficiently sensitive and accurate for SV detection (see
129 Supplemental Results and Fig S3).

130 We then surveyed the distribution of genes and variants. Annotated genes are primarily located
131 towards the telomeres, and most of the identified SNPs are distributed in the gene-sparse regions
132 flanking the centromeres (Figure 1A). In contrast, detected SVs were mainly situated in the gene-
133 rich regions (Figure 1A, B). Frequent translocations and inversions were observed from the
134 breakpoints in gene-rich regions (Figure 1B). Even though the density of called SVs is higher in
135 gene-rich regions, only 0.2% of these SVs affected exons directly.

136 Due to the limitations of the SV detection algorithms based on short reads, the length of the INV
137 and TRA cannot be precisely inferred from the positions of the two breakpoints of an SV. We
138 further examined the length distribution of DEL, DUP, and INS, which showed that most SVs
139 were relatively small, but a substantial minority are large: 30-250 bp: 30.3%; 250-500 bp: 13.1%;

140 500 bp-1 kb: 13.9%; 1 kb-2 kb: 9%; > 2 kb: 33.6%. Two size bands of enrichment were observed
141 around 75 bp and 250 bp for DEL (Supplemental Fig S4). There were also obvious peaks around
142 150 bp and 60 bp for DUP and INS, respectively. These may reflect specific, abundant mobile
143 elements. Sequence composition survey of the SVs indicated that the two most abundant
144 transposable element sequence signatures were *Gypsy* and *EnSpm* (Supplemental Fig S5). These
145 well-known LTR transposable elements play significant roles in plant genome structure and
146 evolution.

147 **The “Domestication Syndrome” in Sorghum: photoperiod sensitivity and variety type**

148 To explore the population structure of the BAP based on SVs, we first investigated the
149 distribution of SVs across the BAP. Structural changes identified in sweet sorghum and typical
150 grain sorghum were fewer than those observed in cellulosic sorghum (Figure 2A, Supplemental
151 Fig S6). Photoperiod sensitivity is a key trait that must be modified to reconcile environmental
152 cues, reproductive cycles and planting/harvest during crop domestication and radiation from
153 center of origin. Modification of photoperiod sensitivity is accompanied by the occurrence of
154 other domestication traits, considered collectively the “domestication syndrome” (Allaby et al.
155 2008; Liu et al. 2015; Song et al. 2017; Lu et al. 2020). Genotype data from a total of 339
156 sorghum lines with variety type and photoperiod information were used for population structural
157 analyses. Principal components analysis (PCA) based on SNP, SV, or combined SNP+SV
158 datasets showed a similar population structure pattern (Figure 2B, Supplemental Fig S7). We
159 examined the deviation regions of the first two principal components ($PC1 > 50$ and $PC2 < -50$)
160 in SV PCA results and found, as expected, that the photoperiod sensitivity feature is strongly
161 linked with cellulosic sorghum while the derived sweet sorghum has photoperiod insensitive
162 characteristics (Figure 2B). Sorghum, unusually, has bidirectional gene flow between

163 wild/weedy relatives and cultivated sorghum lines in sympatric and allopatric species (Mace et al.
164 2013). Exceptions to the population clusters may reflect gene flow. Grain and sweet sorghum are
165 not well differentiated, although SVs show somewhat better separation than SNPs for these
166 variety types. This finding prompted us to explore the relationship between photoperiod
167 sensitivity and sorghum variety types via haplotype network analysis. As shown in Figure 2C,
168 the edges connecting cellulosic sorghum varieties appear to correspond to those for photoperiod
169 sensitivity, while the edges for sweet sorghum correspond to those for photoperiod insensitivity
170 in minimum spanning trees derived from both SNP and SV datasets.

171 **Identification of SVs Underlying the Divergent Evolution of Cellulosic and Sweet Sorghum**

172 Structural sequence divergence, initiating from hotspots along chromosomes and subsequently
173 expanding through the accumulation of minor genomic variants, has been found to be an
174 important driver of divergent evolution (Song et al. 2002). For the purpose of investigating the
175 location of structural genetic differences that may underlie the divergent evolution of cellulosic
176 and sweet sorghum, we curated 43 cellulosic ($PC1 > 50$) and 33 sweet ($PC2 < -50$) sorghum
177 lines from the BAP with consistent genetic clustering based on the SV PCA results (Figure 2B,
178 Supplemental Table S3, S4). Genetic relatedness analyses based on both SNP and SV datasets
179 were performed. The maximum likelihood tree based on the SV dataset shows as expected that
180 the selected cellulosic (solid red pentagram) and sweet (hollow red pentagram) sorghums were
181 each grouped into one cluster (Figure 3A, B). These results indicate that the curated sorghum
182 lines potentially underwent strong variety-specific selections during sorghum domestication and
183 breeding. In order to investigate the fixation index of the SVs between selected cellulosic and
184 sweet sorghum groups, F_{ST} for each site was estimated between cellulosic group and sweet group
185 in the BAP based on whole-genome SNPs. Prior to establishing the selection threshold, we

186 examined the F_{ST} distribution in our study, and found that it captured the top 1% of the SNP F_{ST}
187 distribution when $F_{ST} \geq 0.15$. Hence, we considered $F_{ST} \geq 0.15$ a robust threshold for our
188 selection analysis. There were 1,637 SNPs shown to be highly differentiated between cellulosic
189 and sweet subpopulations with $F_{ST} \geq 0.15$ (from 0.15 to 0.36). SVs between the curated 43
190 cellulosic and 33 sweet sorghum lines were then compared with the loci of the 1,637 highly
191 differentiated SNPs. Comparison showed that 76% (1,250/1,637) of the highly differentiated
192 SNPs were adjacent to at least one SV (range from 1 to 45 SVs) within 10 kb (Supplemental
193 Table S5). This result indicates that the SVs identified between the curated 43 cellulosic and 33
194 sweet sorghum lines likely underwent strong selection while accompanied by the closely linked
195 SNP loci, and the 43 cellulosic and 33 sweet sorghums selected based on the SV PCA results
196 were representative lines that underwent differential selection during the divergent improvement
197 of cellulosic (tropical landraces) and sweet sorghum subpopulations.

198 To find potential hypervariable regions across the groups, we then examined the SV detection
199 frequency in these two groups, consisting of the 43 curated cellulosic and 33 curated sweet
200 sorghum lines respectively, across 1Mb windows. Common SVs that were present in both
201 cellulosic and sweet sorghum groups were excluded to reduce the background noise. Genomic
202 regions with obvious variable SV frequency between the representative cellulosic and sweet
203 sorghum groups were observed (Figure 4A). The heatmap of SV detection frequency manifested
204 that 186 out of 688 SV frequency windows, including 73 continuous genomic regions, showed
205 significant differences (adjusted p value < 0.01 and average SV difference between two groups
206 was ≥ 20) in frequency between representative cellulosic and sweet sorghums (Figure 4B,
207 Supplemental Table S6). Some hotspots of SV frequency we detected have been reported in
208 previous publications. 56 - 57 Mb on Chromosome 1 and 61 - 62 Mb on Chromosome 2 have

209 been identified as hotspots for controlling protein, starch, and amylose content (Ayalew et al.
210 2022). In addition, 52.23 - 61.18 Mb on Chromosome 1, 2.52 - 11.43 Mb on Chromosome 2 and
211 1.32 - 3.95 Mb on Chromosome 3 were also hotspots for source-sink related traits (Chiluwal et al.
212 2022). Boatwright et al. identified 18 genomic regions under selection across six generic
213 sorghum subpopulations underlying the evolutionary divergence during domestication
214 (Boatwright et al. 2022). Six out of ten selection regions with prior QTL information were
215 covered by our identified SV hotspots while only one out of eight selection regions without prior
216 QTL information was covered by our identified SV hotspots.

217 **Structural Variations Reveals Extensive Contributions to Heritability**

218 Decades of studies have provided evidence that, despite their rarity compared to SNPs, SVs
219 account for a substantial fraction of characterized molecular genetic variation with phenotypic
220 consequences (Freeman et al. 2006). To examine the likely impact of the identified SVs on gene
221 function, we evaluated the predicted functional effects of the variants in our SV and SNP
222 datasets. As shown in Supplemental Fig S8, SVs were more likely to have large impacts on gene
223 function, such as duplication, exon loss, codon frame shift and transcript ablation, whereas SNPs
224 generally were predicted to have lower impacts. The annotation of the predicted impacts of SNPs
225 and SVs on sorghum gene function suggested that SVs could have a significant impact on
226 functional genetic variation in sorghum.

227 We then investigated the potential contributions of our SV set to the inheritance of 29
228 quantitative traits and one binary trait (Supplemental Table S7) (Brenton et al. 2016; Brenton et
229 al. 2020). The overall proportion of variance explained by the additive effect of genomic variants
230 (narrow-sense heritability) was estimated by using mixed model analysis for each trait for whole-
231 genome SNP variation only, and then a combined set of SNPs and SVs, i.e. SNP + SV. The

232 estimated heritability ranged from 2% - 57% (median 20%) when we considered only SNP
233 variation. However, the estimated heritability increased substantially, by 16%-99% (median
234 26.5%) for all but one trait (2015_ADF, which stands for acid detergent fiber content in 2015),
235 when taking both SNP and SVs into account. The additive effect of SNP + SV was particularly
236 marked for the trait of photoperiod sensitivity, and for the sorghum variety type itself when used
237 as a phenotype (Figure 5A). Compared with the heritability contributed by SNP data alone, the
238 reduced heritability of 2015_ADF for SNP+SV (from 57% to 33%) likely resulted from the
239 opposite additive effects contributed by SNP and SV datasets separately. Overall, CNV-type
240 variations consistently produced higher heritability estimates than SNPs for nearly all traits, and
241 explained 6.2% more of the phenotypic variance than REA type variations (Figure 5B). These
242 findings show that, though SNPs are generally able to capture the bulk of the heritable genetic
243 effects on phenotype, SVs accounted for a substantial proportion of the missing heritability in
244 SNP-based analysis for most traits.

245 **Structural Variant Data Allows Detection of New GWAS Associations**

246 To further investigate the causative genomic loci associated with the increased heritability gained
247 by adding SVs to the polymorphism dataset, we performed GWAS based on whole-genome SNP,
248 SV, and combined SNP and SV datasets. Firstly, we investigated associations with a sorghum
249 seed pericarp pigmentation trait, “Pericarp_pigmentation”, a well-studied trait whose global
250 variation is due largely to the *Y* locus, which encodes a MYB transcription factor *Yellow seed1*
251 (*Y1*), though the causative variants in this gene have not been definitely identified (Ibraheem et al.
252 2010; Morris et al. 2013b; Rhodes et al. 2014). GWAS based on SV found three significant
253 association signals for seed pericarp pigmentation, including an SV underlying the *Y1* gene
254 (*Sobic.001G397900*) as expected (Figure 6A), while SNP and SNP+SV analyses did not detect

255 association at this locus (Figure 6B, C). The SV (1.5 kb downstream of *YI*) underlying the *YI*
256 locus was called as a translocation from Chromosome 1 to Chromosome 4 (Figure 6D). The
257 breakpoint on Chromosome 4 was also detected by SV-based GWAS. Another substantial SV
258 association signal was detected on Chromosome 8. The polymorphism associated with this locus
259 is a 2.6 kb DEL/INS located 3.2 kb upstream of *TIM22-2* (*Sobic.008G111800*), a mitochondrial
260 import inner membrane translocase and a homolog of a protein involved in seed development in
261 *Arabidopsis* (Zhang et al. 2023b). Further haplotype analyses of the TRA allele underlying the
262 *YI* locus on Chromosome 1 and the 2.6 kb DEL/INS on Chromosome 8 validated their
263 significant correlation with phenotypic variance in “pericarp_pigmentation” (Supplemental Fig
264 S9A-D, S10A-D, see Supplemental Results for details). Our GWAS results for seed pericarp
265 pigmentation based on SVs thus not only found a significant SV association for the well-studied
266 *YI* locus, which was not detected in SNP GWAS, but also identified a potential translocation
267 involved in the genesis of this locus and a compelling new candidate gene for the control of seed
268 pericarp pigmentation.

269 To further confirm the enhanced detectable heritability conferred by SVs in GWAS, we surveyed
270 the number of significant variations detected in GWAS based on each of SV, SNP, and SNP+SV
271 datasets for an additional 29 morphological and physiological traits (Supplemental Table S7)
272 (Brenton et al. 2016; Brenton et al. 2020). We detected the largest number of GWAS
273 associations using the combined SNP+SV dataset, including 234 SV hits and 43 SNPs hits. This
274 was substantially larger than the number of signals (212 hits) detected in SV-alone GWAS. By
275 far the fewest signals were detected in SNP-only GWAS (50 SNP hits). The number of
276 significantly associated loci in SNP-only GWAS was by far the lowest for all traits except days
277 to harvest. SV or SNP+SV found the largest number of significant association signals for all

278 traits (Supplemental Table S10). SNPs hits in SNP+SV GWAS were also observed in SNP-based
279 GWAS for most traits (except for “Total_fresh_weight”: 3/5), with SNP-only GWAS finding
280 more associated SNPs for several traits (likely as a result of an altered multiple-testing
281 correction). Interestingly, however, SVs between SNP+SV and SV-based GWAS results had
282 only 32.6% of loci in common (median across traits: 19.1%) (Figure 7, Supplemental Table S10).
283 This finding indicates that association analysis based on SNPs and SVs separately, as well as the
284 integrated SNP+SV dataset, can each yield distinct and potentially important associations.

285 Considering that sorghum variety type is associated with photoperiod sensitivity (Figure 2B, C
286 and Figure 3), we further investigated the genetic mechanisms that may underlie their divergent
287 evolution by using three GWAS analyses based on SNP, SV, and SNP+SV datasets for six
288 photoperiod sensitivity-related traits, and 23 traits related to the differentiated sorghum variety
289 types (Supplemental Table S7). There were 171 significantly trait-associated SVs detected in SV
290 GWAS, 33 SNPs were detected in SNP GWAS, and 182 variants including 152 SVs and 30
291 SNPs detected in GWAS based on SNP+SV dataset, of which just 21 SVs were common with
292 those from SV GWAS, while all significant SNPs found were in common with those detected
293 using SNP-based GWAS (Supplemental Table S10). In total, 238 polymorphisms, containing
294 228 SVs and 10 SNPs, were identified as significantly associated with the sorghum differentiated
295 variety type-related traits, while 97 variants, including 74 SVs and 23 SNPs, were found to be
296 significantly associated with photoperiod sensitivity traits. There were 65 polymorphisms,
297 including 54 SVs and 11 SNPs, that were associated with at least two different traits. Amongst
298 these variations, we identified a potentially pleiotropic SV associated with multiple traits,
299 *sv_529156_Chr09_59249767*, a 1.3 kb DEL from 59,249,767 bp to 59,252,667 bp on
300 Chromosome 9, located 11.3 kb upstream of a CCT domain-containing gene, *Sobic.009G259100*.

301 Not only is this locus significantly associated with days to harvest (in both 2014 and 2015) and
302 stalk height, but it is also linked with multiple variety type-related traits: “Dry_tons_per_acre”,
303 “Dry_Weight” and “Total_fresh_weight”.

304 Candidate genes within 20 kb of each breakpoint were then investigated for each significant
305 polymorphism. We found 242 candidate genes, such as *dof21*, *SNAC1* and *TEOSINTE*
306 *BRANCHED 1 (tb1)* close to SVs associated with sorghum variety type-related features and 69
307 candidate genes, including likely orthologs of the *Arabidopsis* genes *FL* and *FAR-RED*
308 *ELONGATED HYPOCOTYL 3 (FHY3)* adjacent to SVs associated with photoperiod sensitivity
309 traits (Supplemental Table S11). We noted that certain genes were annotated as potentially
310 involved in agronomic variety traits, but were also associated with the photoperiod sensitivity
311 traits, whilst some known photoperiod related genes were adjacent to SVs associated with usage-
312 related traits. This finding illustrates the relationship between the sorghum usage or variety type
313 and photoperiod sensitivity; for example, modern grain or sweet sorghum varieties will be
314 expected to flower at different latitudes and times than forage or biomass sorghum. Based on
315 analysis of all traits, we selected 13 candidate loci which were correlated with both photoperiod
316 sensitivity and sorghum variety usage type (Supplemental Table S12).

317 **SVs Have Widespread Impacts on Gene Expression**

318 By modifying the sequence or location of *cis*-regulatory elements, splicing of a gene, copy
319 number, or regulatory RNA molecules, SVs can readily alter the expression pattern of genes (Li
320 et al. 2012; Alaei-Mahabadi et al. 2016; Chiang et al. 2017; Alonge et al. 2020). To explore the
321 impact of SVs on gene expression, we performed RNA sequencing (RNA-seq) on 4 sorghum
322 inbred lines included in the BAP: BTx623, which is a typical grain sorghum and also used as the
323 standard reference genome in our study, RTx430, a grain sorghum inbred with a repeat-rich

324 genome (Deschamps et al. 2018), and Tracy and Ramada, which are typical sweet sorghum lines.
325 Gene expression profiles were generated for both leaf and stem, at 3 stages: pre-flowering,
326 flowering, and milk. Due to the limitation of associating other types of SV with specific genes,
327 only CNV-type variations (DEL, DUP, INS) were taken into consideration for this analysis.
328 Hypergeometric testing was used for enrichment analysis of differentially expressed genes
329 (DEGs) in SV-associated genes. The p values were adjusted using the Bonferroni correction.
330 More DEGs were associated with SVs than not. The percentage of SV-associated DEGs were
331 notably higher than the percentage of non-SV-associated DEGs across all tissues and
332 developmental stages, and the DEGs were significantly enriched in the SV-associated genes
333 (Figure 8A, Supplemental Fig S11A, Supplemental Table S13). SVs with higher predicted
334 impact on the sorghum genome were associated with more DEGs than the SVs with lower
335 predicted impact (Figure 8B, Supplemental Fig S11B); however, the percentage of the DEGs
336 associated with the lower predicted impact SVs were still higher than the percentage of non-SV-
337 associated DEGs: with an average of 9.31% vs 3.21% in leaves and 9.57% vs 4.96% in stems
338 across different accessions and development stages (Supplemental Table S13). Some previously
339 reported genes of phenotypic interest were found among the identified SV-associated DEG set,
340 such as the *Dry* gene, which is an important gene controlling the stem pithy/juicy trait (Zhang et
341 al. 2018), *SUT5*, which encodes a sucrose transporter (Cooper et al. 2019), *Heading Date 1* (Liu
342 et al. 2015), *lipid-transfer protein 1* (Pelèse-Siebenbourg et al. 1994), *gs*, which is a glutamine
343 synthetase gene affects growth and development in sorghum (Urriola and Rathore 2015), and
344 *Ae1*, which is associated with grain quality in sorghum (Figueiredo et al. 2010). These findings
345 suggest that SVs are strongly associated with heritable differential gene expression across

346 varieties, giving a plausible mechanism by which SVs may have a disproportionate impact on
347 phenotypic variation.

348 **Discussion**

349 Recent studies have revealed an abundance of large-scale genomic variants in many plant species,
350 but the effects of SV on global variation of quantitative traits are not yet established. Here we
351 built an SV dataset based on Illumina whole-genome data for 363 sorghum lines. The apparent
352 discrepancy between detected SNP and SV distribution in the genome (Figure 1A, B) may
353 illustrate the different mechanisms of creation and mutation of SVs and SNPs. We examined in
354 detail the representative 43 cellulosic and 33 sweet sorghum lines from the BAP. Structural
355 genetic differences underlying the divergent evolution of the representative cellulosic and sweet
356 sorghum lines helped us demonstrate the extent of the role played by SVs in sorghum variety
357 type differentiation, and provide potential targets for sorghum breeding and engineering. GWAS
358 based on whole-genome structural variation revealed novel genetic associations and new
359 candidate genes for sorghum seed pericarp pigmentation, which were not detected in previous
360 GWAS or our SNP-alone analysis. Strong and extensive correlations between SVs and sorghum
361 phenotypes were observed in subsequent association analysis for 29 additional traits. For most of
362 these traits, heritability was improved by the addition of SVs to the extensive set of SNPs, in
363 some cases substantially so, and in many cases, associations were detected that were not seen in
364 SNP data alone. RNA-seq analysis of four sorghum lines in two tissues and three developmental
365 stages demonstrated impacts of SVs on gene expression in the sorghum genome. These findings
366 show that the SV dataset we built is a powerful addition to GWAS analysis in sorghum,
367 providing insights into key loci underlying sorghum adaptation and improvement, mechanisms

368 of variation in gene expression, and improved methodologies to maximize discovery of causative
369 genetic alleles.

370 Limitations in sensitivity and specificity are perhaps the main reason why SV analysis has not
371 yet been more widely used in crop genetics. There are three strategies partly or completely
372 applied to SV calling in current popular algorithms for short-read sequencing datasets: read-pair
373 technologies, read-depth methods, and split-read approaches, all of which are based on aligning
374 sequencing reads to a reference genome and detecting discordances underlying the SVs (Alkan et
375 al. 2011). Depending on the type of variants or the features of the underlying sequence at the SV
376 locus, each algorithm has different strengths and disadvantages in terms of SV detection. The
377 weaknesses can be overcome to some extent by extracting the consensus of multiple algorithms
378 based on different strategies in an ensemble approach, as applied here (Zarate et al. 2020). The
379 common limitation of the short read reference-based structural variant callers is that they are
380 heavily biased against insertions relative to the reference, unsurprisingly since inserted sequences
381 do not appear in the reference genome (The 1000 Genomes Project Consortium 2010; Mills et al.
382 2011). Long-read sequencing technologies and assembly-based methods are therefore necessary
383 to provide complete coverage of SVs, particularly insertional polymorphisms, and to fully
384 understand the sequence underlying the different allelic forms of SV. We show here that the
385 ensemble approach, while necessarily incomplete, is nonetheless a powerful addition to
386 understanding causative genetic variation; pangenome construction using long read technologies
387 will further validate our results and help complete the SV datasets in the future.

388 As a comparator for the short-read based methods, we used whole-genome alignment of
389 assemblies based on long-read technologies. The MUMmer system, and the genome sequence
390 aligner NUCmer included within it, have been widely used for alignment at genome scale

391 (Marcais et al. 2018). Many approaches for variant calling by assembly comparison use the
392 MUMmer system for the genome-scale alignment step. In this study, we used MUM&Co (v3.7)
393 (O'Donnell and Fischer 2020) to call the SVs from five genome assemblies against BTx623 to
394 provide a ground truth in order to evaluate the SV calling approach we deployed. A substantial
395 number of SVs were identified by the mate-pair based fusion pipeline that did not have a clear
396 match with any of the SVs called by MUM&Co. To cross-reference the accuracy and validity of
397 MUM&Co, we compared the SVs datasets called by MUM&Co to those identified by
398 Assemblytics (Nattestad and Schatz 2016), which is also derived from the MUMmer system. The
399 interpretation of complex structural variations posed challenges for these evolving whole-
400 genome comparison methods. We found substantial discrepancies even between the SVs datasets
401 called by MUM&Co and by Assemblytics for the five genomes we compared to the BTx623
402 reference. The SVs called by Assemblytics also heavily depend on the “unique sequence anchor”
403 and “maximum variant size” parameters, while MUM&Co can produce very large artifactual
404 SVs, again making the maximum size threshold a critical parameter. Even when they utilize the
405 same widely accepted aligner, the inconsistency and parameter sensitivity of whole-genome
406 comparison methods limit their utility, especially for larger variants. We therefore conclude that
407 short-read methods remain a valid and cost-effective approach for SV detection, with no decisive
408 disadvantages when using a single reference approach.

409 **SV GWAS and heritability**

410 Importantly, by adding structural variant data to GWAS analysis, we found additional significant
411 association peaks. In other words, SNPs alone do not identify all the detectable LD blocks in
412 association with the target traits. This violates the basic assumptions of GWAS (Lipka et al.
413 2015), because genome-wide SNP data should provide multiple polymorphisms within the range

414 of LD for each causative locus, even if the causative locus is an SV not detected by SNP
415 genotyping. However, recent studies have shown that SVs causing important trait variation in
416 crops (for example, soybean protein and oil content (Fliege et al. 2022)) are not always in strong
417 LD with surrounding SNPs, because of transposon excision, illegitimate recombination, and
418 other mechanisms independent of the Mendelian assumptions underlying LD calculations.
419 Notably, by including SVs in our GWAS, we were not only able to identify more loci in
420 significant association with traits, but we also substantially increased the measured narrow-sense
421 heritability for some traits, in one case approaching the maximum value of 1. Previous studies in
422 other species have also shown the power of SVs to identify missing heritability (Jeffares et al.
423 2017; Alonge et al. 2020). The capacity of SVs to capture missing heritability could be attributed,
424 at least in part, to their frequent direct impact on gene expression. Chiang et al. performed the
425 eQTLs mapping using joint analysis of SVs, SNVs, and indels in human, and observed a notable
426 abundance of SV-associated gene expression (Chiang et al. 2017). Our findings confirm in
427 sorghum that missing heritability may be at least partially due to SVs that are not in strong LD
428 with any local SNPs.

429 **Potential for SV-driven breeding of sorghum**

430 Sorghum is a good resource for bioenergy production, and production of lipids is of increasing
431 interest to remedy the world-wide energy crisis (Sandesh and Ujwal 2021). To verify the
432 possibility of sorghum as a feedstock for oil production by SV-driven breeding, we identified
433 331 orthologs characterized as involving oil synthesis in *Arabidopsis* (Supplemental Table S14).
434 We predicted the potential functional effects of SVs on the oil-related genes. 96% (323/331) and
435 99% (328/331) of the oil gene orthologs were associated with CNV type SVs and rearrangement
436 type SVs, respectively (Supplemental Table S15, S16). Almost half of the orthologs (48%,

437 159/331) are predicted to be highly impacted by CNV type SVs (Supplemental Fig S12). We
438 found also that DEGs are strongly associated with SVs, even in populations outside the BAP
439 (Supplemental Fig S13-S15, Supplemental Table S17, see Supplementary Results for details)
440 These results suggest that bioenergy traits, including oil traits, could be enhanced via breeding
441 endeavors, and that specific targeting of SVs via marker-assisted selection could allow
442 modification of gene expression levels in many cases.

443 Altogether, our study highlights the complementary contribution of the underexplored SVs in
444 heritability of important traits, reveals their widespread impacts on gene expression, and
445 demonstrates their crucial role in shaping population genetic diversity as well as trait
446 determination. The findings in our study have significant implications for crop breeding and
447 improvement, underscoring the indispensable role of SVs in future studies.

448 **Methods**

449 **Re-sequencing dataset and phenotypes**

450 The Illumina short-read sequence dataset and phenotypes of the sorghum lines used in this study were collected by the TERRA-
451 REF project <http://terraref.org> (Brenton et al. 2016); 339 sorghum lines with population information were considered for
452 population genetic analysis. Information for each line is included in Supplemental Table S1.

453 **Plant tissue and sequencing**

454 Leaves from the seedlings of sorghum were sampled in the greenhouse. At least 10g of leaf tissue for each sorghum accession
455 was sent to Roy J. Carver Biotechnology Center at the University of Illinois at Urbana-Champaign. Raw HIFI sequence data in
456 BAM format was generated by PacBio Sequel IIf platform.

457 **Variant calling**

458 SNPs were called using the Sentieon (version 202010.01) (Kendig et al. 2019) DNA-seq pipeline. Ensemble variant calling using
459 five independent tools based on different algorithms was used to call structural variations (SVs). See Supplemental Methods for
460 details.

461 **De novo assembly and comparison**

462 BAM files were converted to FASTQ files by SAMtools (Li et al. 2009). Reads less than 1 kb were identified and filtered by
463 SeqKit tools (Shen et al. 2016). Genome *de novo* assembly were performed by hifiasm (Cheng et al. 2021). Genome assembly
464 quality was evaluated by quast (Gurevich et al. 2013), and BUSCO (Simao et al. 2015). MUM&Co(v3.7) (O'Donnell and Fischer
465 2020) was used to evaluate SVs based on assembly comparison.

466 **The heatmap of SV detection frequency**

467 The heatmap of SV detection frequency was built individually in 1Mbp sliding windows for the representative 43 cellulosic and
468 33 sweet sorghum lines to identify regions with elevated genetic differentiation. In order to reduce the noise from the background,
469 SVs that were present in both cellulosic sorghum lines and sweet sorghum lines were excluded individually. The *p* values for the
470 difference tests were adjusted using Bonferroni correction, significance hypervariable regions were defined as adjusted *p* value <
471 0.01 and average SV difference between two groups was ≥ 20 .

472 **Heritability estimation**
473 LDAK (v5.1) (Zhang et al. 2021) was used to estimate the trait heritability explained by the SNP and SV polymorphisms.

474 **Population genetics analysis**
475 *SNPRelate* (Zheng et al. 2012). SVs were converted to present-absent binary representation before conducting PCA. F_{ST} was
476 calculated by using VCFtools (v0.1.16) (Danecek et al. 2011). Principal component analysis was performed using the R function
477 *prcomp()* (R core Team, 2022). A minimum spanning tree was created using the R package *Poppr* (Kamvar et al. 2014). SNPhylo
478 (Lee et al. 2014) was used to create maximum likelihood phylogenetic trees.

479 **GWAS**
480 GWAS was performed by GAPIT3 using the compressed mixed linear model (CMLM) model (Zhang et al. 2010; Wang and
481 Zhang 2021). See Supplemental Methods for details on SV association methods.

482 **Haplotype analyses**
483 The R package *geneHapR* (Zhang et al. 2023a) was used to perform analyses.

484 **RNA-seq analysis**
485 Tissues samples for RNA were collected from plants grown in the field at the Energy Farm at the University of Illinois at
486 Urbana-Champaign in 2018. RNA-seq data were analyzed using the DESeq2 package (Love et al. 2014), and plots drawn by
487 ggplot2 (Villanueva and Chen 2019).

488

489

490 **Data access**

491 The raw sequencing data for the 363 TERRA-REF lines are available at
492 https://datacommons.cyverse.org/browse/iplant/home/shared/terraref/genomics/raw_data/bap/resequencing. The raw gene
493 expression data are available at https://genome.jgi.doe.gov/portal/SorbicEProfiling_31_FD/SorbicEProfiling_31_FD.info.html
494 and https://genome.jgi.doe.gov/portal/SorbicEProfiling_30_FD/SorbicEProfiling_30_FD.info.html. The SV and SNP datasets
495 used in this study are attached in Supplemental_SNP_dataset.vcf.gz and Supplemental_SV_dataset.vcf.gz.

496 **Competing interests**

497 The authors declare no competing interests.

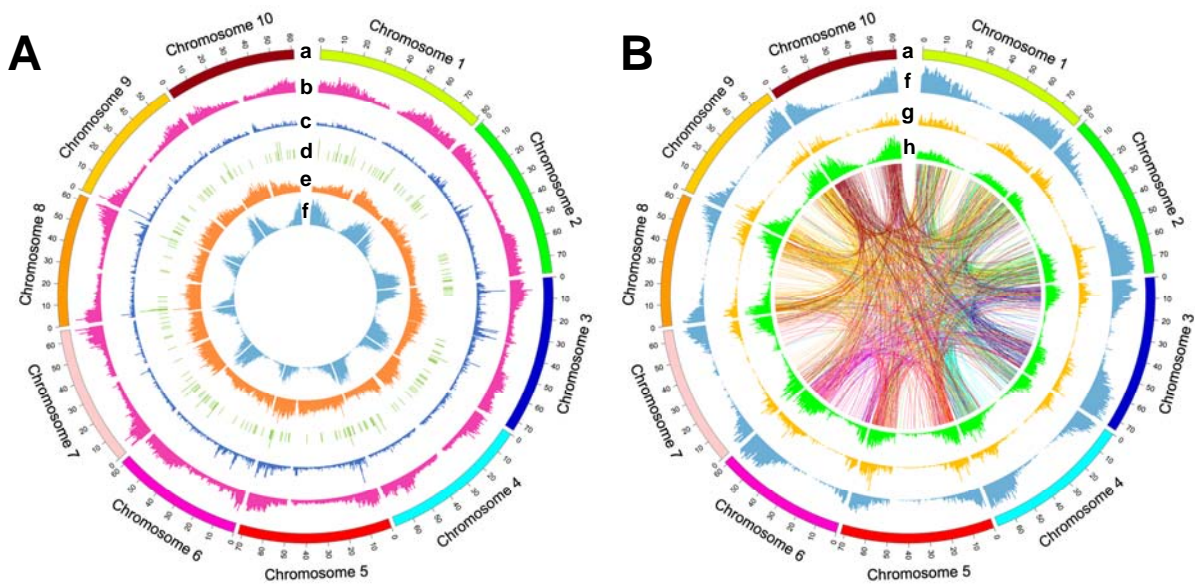
498 **Acknowledgements**

499 We especially wish to express our sincere gratitude to Dr. Amy Marshall-Colon for her
500 assistance with the data storage and computation resources, Drs. Todd Mockler and Jeremy
501 Schmutz for assistance with data access, and Drs. John Vogel and Peggy Lemaux for pre-
502 publication access to the RTx430 genome information. This work was funded by the DOE
503 Center for Advanced Bioenergy and Bioproducts Innovation (U.S. Department of Energy, Office
504 of Science, Office of Biological and Environmental Research under Award Number DE-
505 SC0018420). Any opinions, findings, and conclusions or recommendations expressed in this
506 publication are those of the author(s) and do not necessarily reflect the views of the U.S.
507 Department of Energy. The work (proposal: 10.46936/10.25585/60001277) conducted by the
508 U.S. Department of Energy Joint Genome Institute (<https://ror.org/04xm1d337>), a DOE Office of
509 Science User Facility, is supported by the Office of Science of the U.S. Department of Energy
510 operated under Contract No. DE-AC02-05CH11231.

511 **Author contributions**

512 Dr. Zhihai Zhang helped design the study, performed the analysis, and wrote the manuscript. Dr.
513 Joao Paulo Gomes Viana offered R scripts recommendations. Dr. Bosen Zhang contributed to
514 the protocols of RNA extraction and purification and quantification and samples collection. Dr.
515 Kimberly K.O. Walden offered advice for genome *de novo* assembly. Dr. Hans Müller Paul
516 provided suggestions for Python scripts. Dr. Stephen Patrick Moose contributed to samples
517 collection and RNA samples submission to JGI. Dr. Geoffrey P. Morris assisted with sorghum
518 genetics and phenotype information and provided the phenotypic datasets of BAP. Dr Matthew E.
519 Hudson obtained funding, designed the study, assisted with the analysis, and edited the
520 manuscript.

521



522

523

524 **Figure 1** Distribution of genome-wide variations in the sorghum Bioenergy Association Panel (BAP). **A** Distribution of gene
525 density and copy number variant (CNV) type structural variants (SVs), including deletions (DEL), duplications (DUP) and
526 insertions (INS). From the outermost layer to the innermost layer of the Circos plot represents chromosomes (a), DEL density (b),
527 DUP density (c), INS density (d), single nucleotide polymorphism (SNP) density (e), and gene density (f) respectively.
528 Annotated genes were primarily located flanking centromeres as expected. Most of the identified SNPs were distributed in the
529 gene-sparse regions. CNV-type SVs showed a different distribution pattern than SNPs, and were mainly situated in the gene-rich
530 regions. The densities of genes and CNV-type SVs were calculated in 500 kb windows. **B** Distribution of gene density and
531 rearrangement (REA) type SVs, including inversions (INV) and translocations (TRA). From the outermost layer to the innermost
532 layer of the Circos plot represents chromosomes (a), gene density (f), INV density (g), and TRA density (h) respectively. The
533 core of the Circos plot is a spanning diagram of the identified translocations. The links show the two breakpoints located in
534 different chromosome positions for each TRA. Each link is colored by the chromosome color of the start position of the
535 corresponding TRA. As with CNV-type variations, identified INVs and TRAs were distributed mainly in gene enriched zones.
536 Frequent rearrangement flows were observed between chromosomes. The densities of genes and REA-type variants were
537 calculated in 500 kb windows. The link diagram was evenly thinned (1/256) from the total TRAs.

538

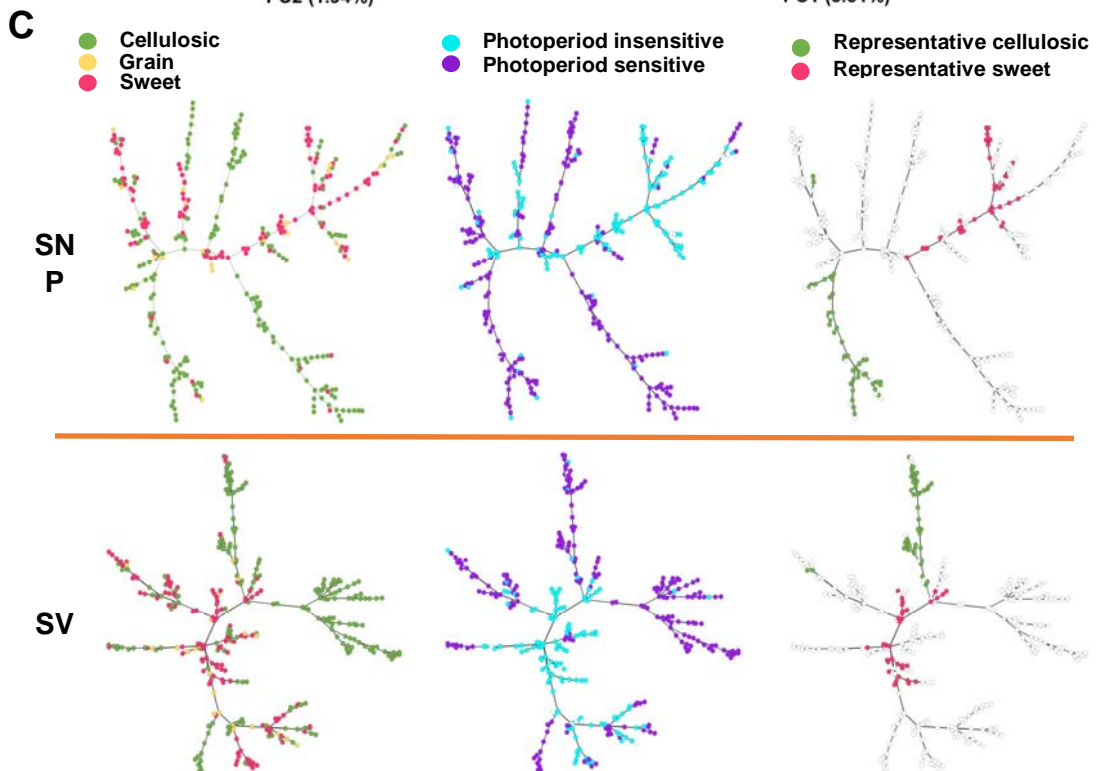
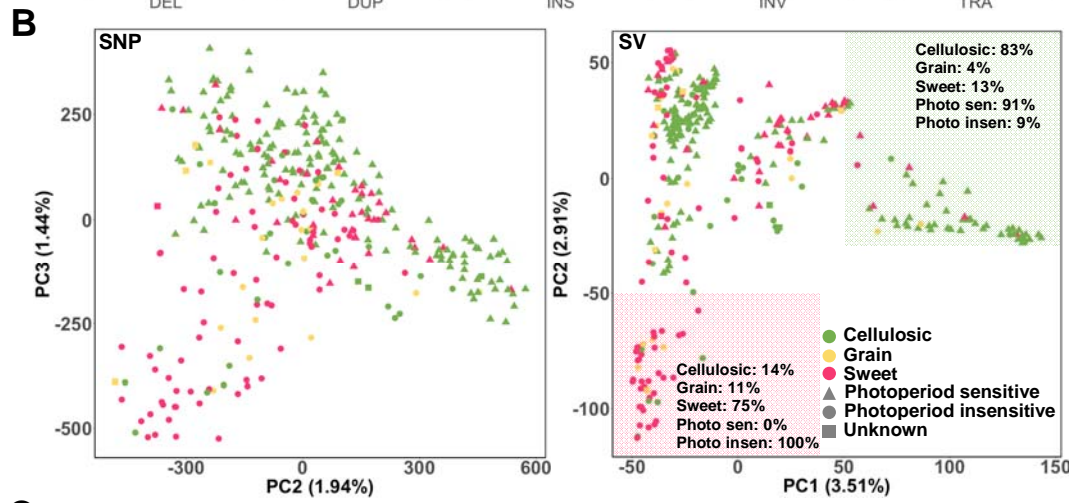
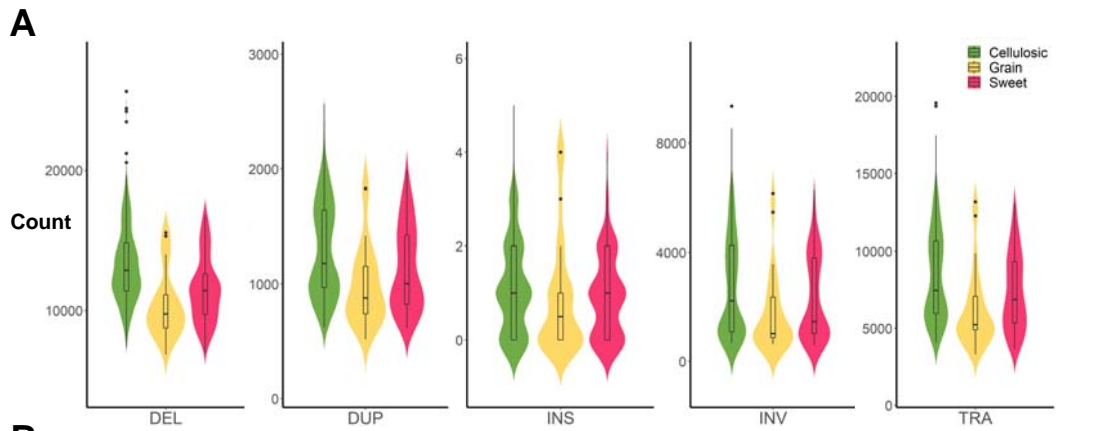
539

540

541

542

543



545 **Figure 2** Structural variation (SV) distributions in different sorghum variety types and population structural analyses. **A** Violin
546 and boxplot for SVs count distributions in cellulosic, grain and sweet sorghum groups. Deletion (DEL), duplication (DUP),
547 insertion (INS), inversion (INV) and translocation (TRA) count distributions were calculated separately in cellulosic (green, left),
548 grain (yellow, center) and sweet (red, right) sorghum groups. Compared with the other two sorghum variety types, cellulosic
549 sorghum contained the most called SVs, indicating that sweet sorghum may be closer to grain sorghum than cellulosic sorghum
550 in SV content as the reference BTx623 is a typical grain sorghum. **B** Principal Components Analysis (PCA) based on single
551 nucleotide polymorphism (SNP, left) and SVs (right). Photoperiod sensitivity: Photoperiod_Insensitive (circle),
552 Photoperiod_Sensitive (triangle) and unknown (square), and sorghum variety type information: cellulosic (grass green), grain
553 (yellow) and sweet (red) were differentiated by PCA based on SNPs and SVs. In SV PCA, the corner in the upper antidiagonal
554 with translucent green background shows the zones with $PC1 > 50$; the corner in the lower antidiagonal with translucent red
555 background shows the area with $PC2 < -50$. The percentages in both colored corners represent the proportions of different
556 sorghums with the corresponding attributes. **C** Minimum spanning trees. Minimum spanning trees were exhibited based on both
557 single nucleotide polymorphisms (SNPs) (top) and SVs (bottom). In the first column, sorghum variety type information is coded:
558 cellulosic (green), grain (yellow) and sweet (red); in the second, photoperiod sensitivity: photoperiod insensitive (sky blue) and
559 photoperiod sensitive (purple). In the third column, distribution of the selected representative cellulosic (grass green) and
560 representative sweet (red) are shown from the PCA analysis. In general, sweet sorghum spreading branches matched those of
561 photoperiod sensitive sorghum lines, while cellulosic sorghum spreading branches matched those of photoperiod insensitive
562 sorghum lines. and variety type.

563

564

565

566

567

568

569

570

571

572

573

574

575

576

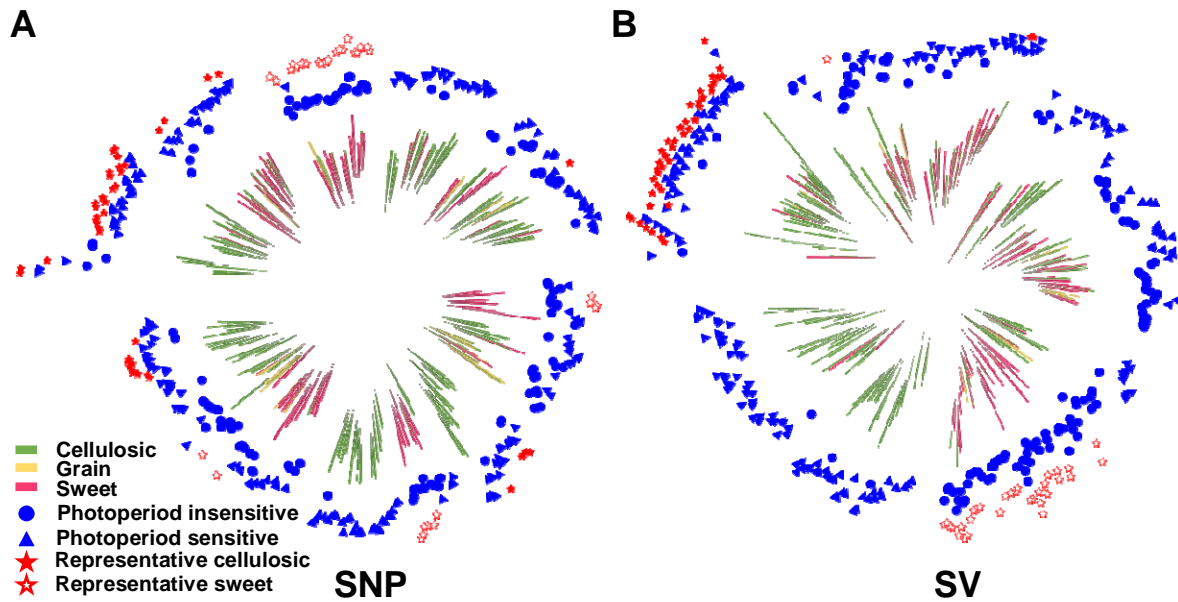
577

578

579

580

581



582

583 **Figure 3** Phylogenetic trees of 339 sorghum lines in the Bioenergy Association Panel (BAP). Phylogenetic trees were conducted
 584 using single nucleotide polymorphisms (SNPs) (**A**) and structural variants (SVs) (**B**) as characters. Sorghum variety type and
 585 photoperiod sensitivity were marked as different colors and shapes: cellulosic (green line), grain (yellow line), sweet (red line),
 586 photoperiod insensitive (blue solid circle), photoperiod sensitive (blue solid triangle), selected representative cellulosic
 587 accessions (red solid pentagram) and selected representative sweet accessions (red hollow pentagram). The maximum likelihood
 588 phylogenetic tree based on the SV dataset shows a clearer classification of phylogeny, sorghum variety type and photoperiod
 589 sensitivity than the maximum likelihood phylogenetic tree based on SNPs, with selected cellulosic and sweet sorghums being
 590 almost monophyletic based on SV data.

591

592

593

594

595

596

597

598

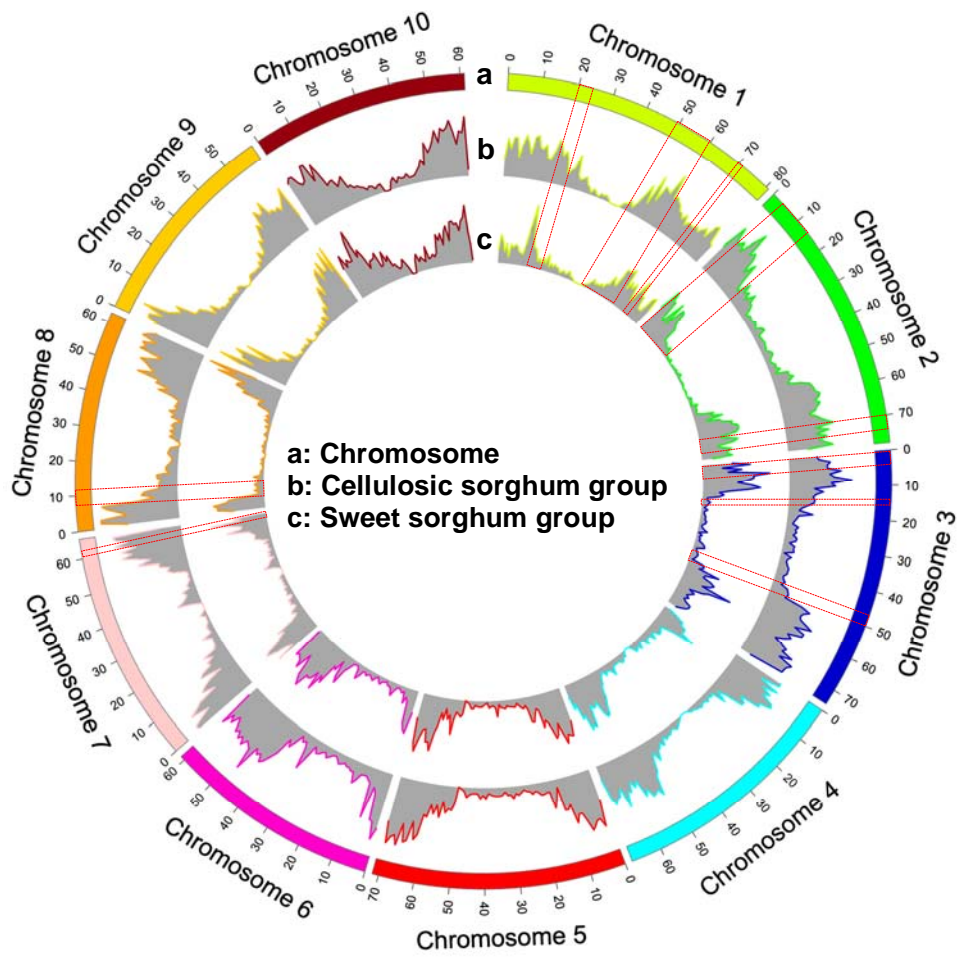
599

600

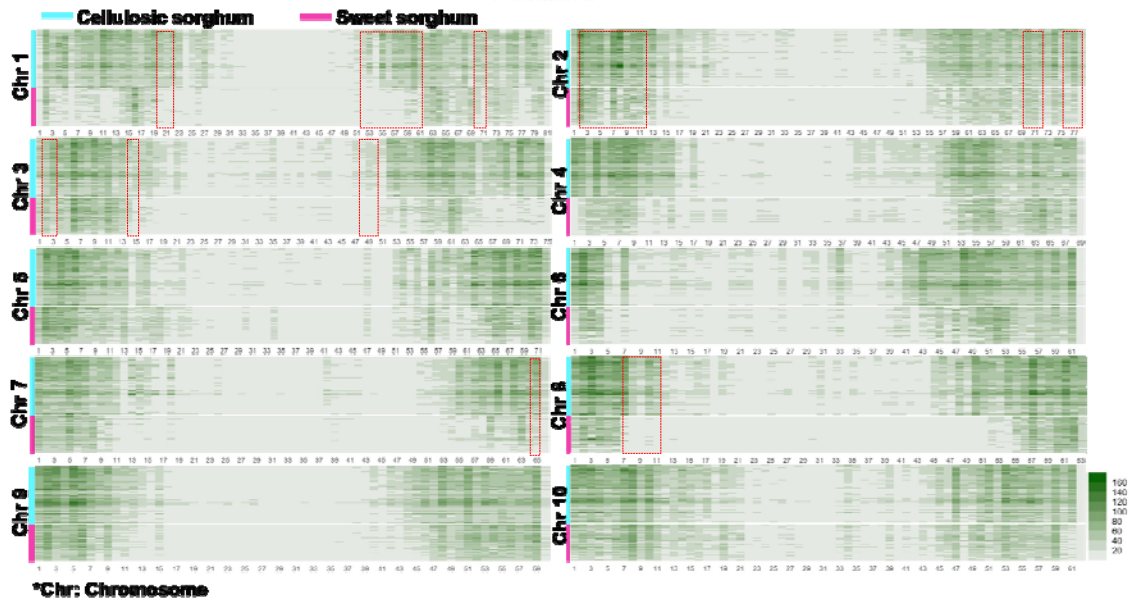
601

602

A



B



603

604

605 **Figure 4** Typical structural variations (SVs) in the divergent evolution of cellulosic and sweet sorghum. **A** Circos plot for the SV
606 frequency differences between the selected representative cellulosic sorghum group and the sweet sorghum group. a,
607 chromosomes. b, SV frequency of cellulosic group. c, SV frequency of sweet group. SV frequencies were calculated in 1MB
608 sliding windows in each group. Hypervariable genomic regions were observed between representative cellulosic and sweet
609 sorghum groups. **B** Heatmap of SV frequency for selected representative cellulosic and sweet sorghum lines. SV frequencies
610 were detected individually and chromosome by chromosome in 1Mb sliding windows. The vertical axis stands for the stacked
611 heatmaps for each sorghum line per chromosome. Green bar showed the range of the stacked heatmaps for cellulosic sorghum
612 lines in each chromosome. Red bar showed the range of the stacked heatmaps for sweet sorghum lines in each chromosome. X
613 axis stands for the physical distance for every chromosome. High SV detection frequencies were observed towards the telomeres
614 in each chromosome for both cellulosic group and sweet group. 186 out of 688 SVs frequency windows were tested as significant
615 difference windows between representative cellulosic and sweet sorghum accessions. Red dash box indicates the hotspots
616 previously reported covered by the 186 significant difference windows.

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

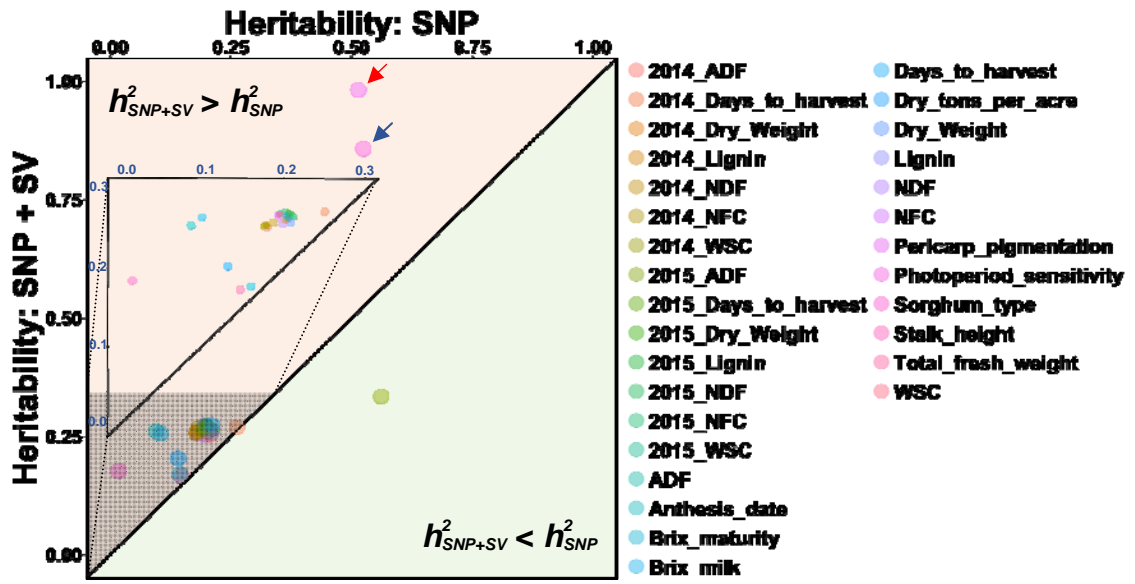
636

637

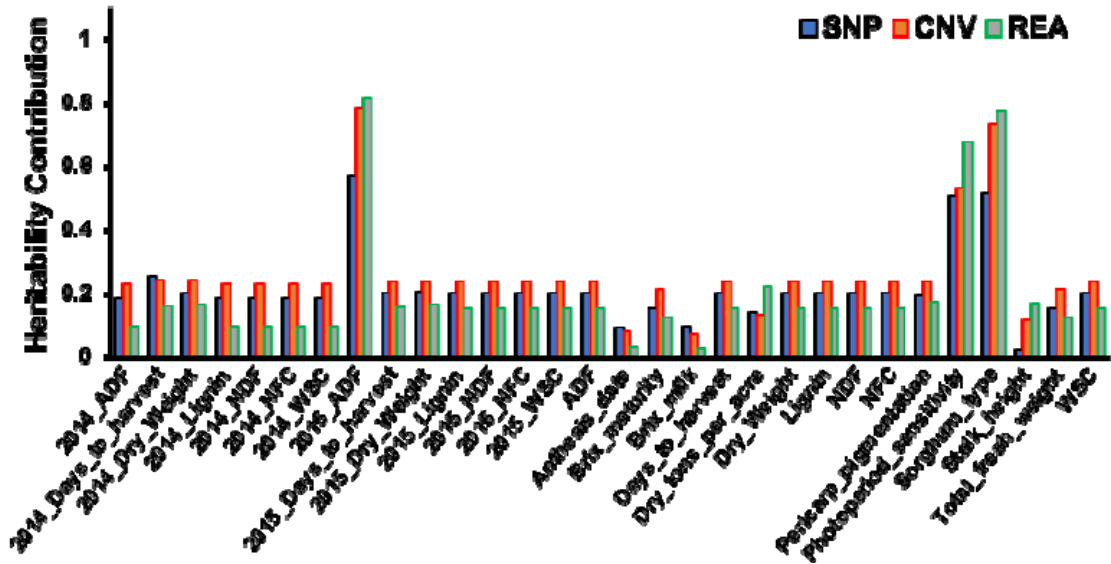
638

639

A



B



640

641

642

643

644

645

646

647

648 **Figure 5** Structural variation (SV) contributes substantially to heritability. **A** Heritability estimates are improved by the addition
649 of SVs. Narrow-sense heritability were estimated for 29 quantitative traits and one binary trait. The upper diagonal colored by
650 melon is the area in which the heritability of single nucleotide polymorphism (SNP) + SV is greater than the heritability of SNP
651 only ($h_{\text{SNP+SV}}^2 > h_{\text{SNP}}^2$). The lower diagonal colored by spring green is the area in which the heritability of SNP + SV is less than the
652 heritability of SNP only ($h_{\text{SNP+SV}}^2 < h_{\text{SNP}}^2$). The diagonal line illustrates where heritability estimates with and without SVs are
653 the same ($h_{\text{SNP+SV}}^2 = h_{\text{SNP}}^2$). 30 traits were dotted by different colors in the plot. The embedded upper triangular dot plot shows the
654 magnification of the shaded area. All of traits, except for 2015_ADF, were observed in the upper $h_{\text{SNP+SV}}^2 > h_{\text{SNP}}^2$ area, which
655 indicates the predicted total heritability increase for most traits when taking both SNP and SVs into account compared with
656 taking SNPs only into consideration. This was particularly marked for two traits: photoperiod sensitivity (pointed by blue arrow)
657 and sorghum variety type (pointed by red arrow). **B** A bar plot for estimation of heritability contribution from SNP, copy number
658 variations (CNV) and rearrangement (REA) type variation. Narrow-sense heritability was estimated for 29 quantitative traits and
659 one binary trait (Supplemental Table S7): "2014_ADF", acid detergent fiber content in 2014; "2014_Days_to_harvest", days to
660 harvest in 2014; "2014_Dry_Weight", dry weight of biomass in 2014; "2014_Lignin", lignin content in 2014; "2014_NDF",
661 neutral detergent fiber in 2014; "2014_NFC", non-fibrous carbohydrates content in 2014; "2014_WSC", water-soluble
662 carbohydrates content in 2014; "2015_ADF", acid detergent fiber content in 2015; "2015_Days_to_harvest", days to harvest in
663 2015; "2015_Dry_Weight", dry weight of biomass in 2015; "2015_Lignin", lignin content in 2015; "2015_NDF", neutral
664 detergent fiber in 2015; "2015_NFC", non-fibrous carbohydrates content in 2015; "2015_WSC", water-soluble carbohydrates
665 content in 2015; "ADF", average acid detergent fiber content of 2014 and 2015; "Anthesis_date", date of anthesis;
666 "Brix_maturity", brix content in maturity stage; "Brix_milk", brix content in milk stage; "Days_to_harvest", average days to
667 harvest of 2014 and 2015; "Dry_tons_per_acre", dry tons per acre; "Dry_Weight", dry weight of biomass; "Lignin", lignin
668 content; "NDF", average neutral detergent fiber content of 2014 and 2015; "NFC", average non-fibrous carbohydrates content of
669 2014 and 2015; "Pericarp_pigmentation", pericarp pigmentation; "Photoperiod_sensitivity", photoperiod sensitivity;
670 "Sorghum_type", sorghum variety type: sweet, grain and cellulosic; "Stalk_height", stalk height; "Total_fresh_weight", total
671 fresh weight; "WSC", average water-soluble carbohydrates content of 2014 and 2015. Blue bars, orange bars and gray bars with
672 green frame stand for the heritability contributions from SNP, CNV-type variations, and REA-type variations respectively. For
673 most of the traits, CNV-type variations explained more variance than REA-type variations.

674

675

676

677

678

679

680

681

682

683

684

685

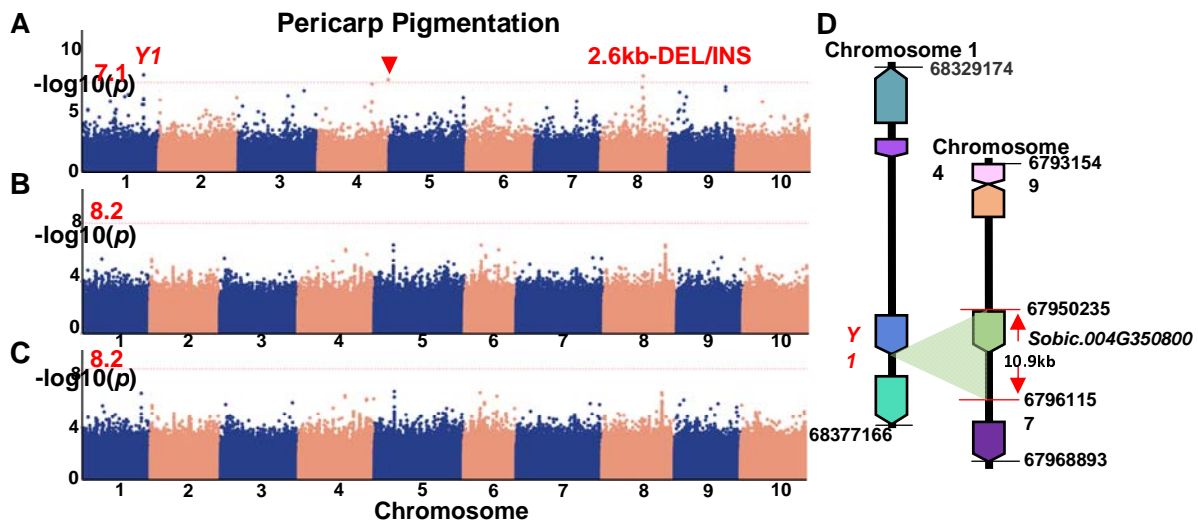
686

687

688

689

690



691

692 **Figure 6** Manhattan plots of genome-wide association study (GWAS) results for “Pericarp Pigmentation”. **A** GWAS result for
 693 the pericarp pigmentation trait based on structural variants (SVs) alone. Three significant signals were detected using compressed
 694 mixed linear model (CMLM) including a signal underlying the well-known pericarp pigmentation related *Y1* gene. The
 695 corresponding signal underlying the *Y1* was a translocation variation between Chromosome 1 and Chromosome 4. The signal at
 696 the other breakpoint on Chromosome 4 of the translocation underlying *Y1* was also detected (solid red inverted triangle). The
 697 signal on Chromosome 8 was a 2.6 kb deletion / insertion (DEL/INS) located near *MYB5* (*Sobic.008G112200*). **B**, **C** GWAS
 698 results for the “pericarp_pigmentation” based on single nucleotide polymorphisms (SNPs) alone (**B**) and SNPs+SVs (**C**). The red
 699 dotted lines in the Manhattan plots show the Bonferroni corrected threshold of $\alpha=0.05$. The red numbers near the red dotted lines
 700 were the corresponding values of the Bonferroni corrected threshold of $\alpha=0.05$ based on different datasets; no loci reached the
 701 corrected significance threshold. **D** A diagram for the translocation underlying *Y1*. The corresponding signal underlying *Y1* was a
 702 translocation with approximately 10.9 kb span including a coding gene (*Sobic.004G350800*) located on Chromosome 4 in the
 703 reference genome.

704

705

706

707

708

709

710

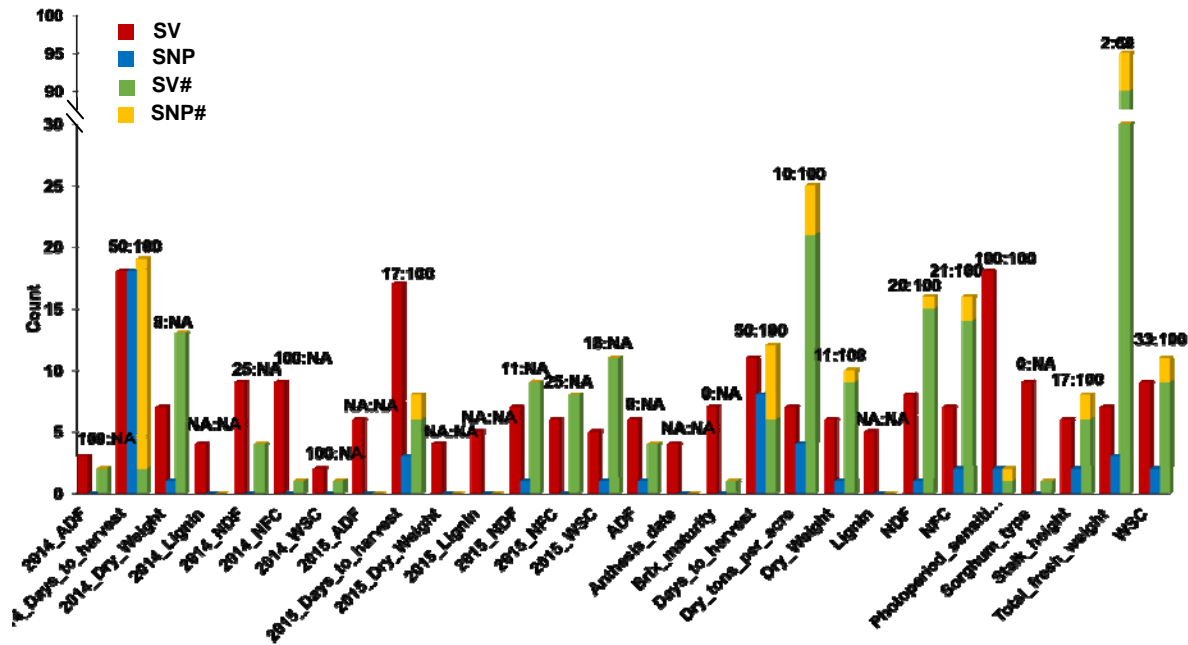
711

712

713

714

715



716

717 **Figure 7** Number of significant genotype-phenotype associations detected in genome-wide association studies (GWAS). Number
 718 of significant associations for 28 traits (Supplemental Table S10, there were 29 traits in total being analyzed, but there was no
 719 significant signal detected for “Brix_milk”) detected in GWAS based on structural variation (SV) (red columns, the first column
 720 per three column set), single nucleotide polymorphism (SNP) (blue columns, the second column per three column set) and
 721 SNP+SV (the third stacked column per three column set, including both SVs (SV#, green), and SNPs (SNP#, orange)) datasets
 722 was showed in the column chart. “2014_ADF”, acid detergent fiber content in 2014; “2014_Days_to_harvest”, days to harvest in
 723 2014; “2014_Dry_Weight”, dry weight of biomass in 2014; “2014_Lignin”, lignin content in 2014; “2014_NDF”, neutral
 724 detergent fiber in 2014; “2014_NFC”, non-fibrous carbohydrates content in 2014; “2014_WSC”, water-soluble carbohydrates
 725 content in 2014; “2015_ADF”, acid detergent fiber content in 2015; “2015_Days_to_harvest”, days to harvest in 2015;
 726 “2015_Dry_Weight”, dry weight of biomass in 2015; “2015_Lignin”, lignin content in 2015; “2015_NDF”, neutral detergent
 727 fiber in 2015; “2015_NFC”, non-fibrous carbohydrates content in 2015; “2015_WSC”, water-soluble carbohydrates content in
 728 2015; “ADF”, average acid detergent fiber content of 2014 and 2015; “Anthesis_date”, date of anthesis; “Brix_maturity”, brix
 729 content in maturity stage; “Days_to_harvest”, average days to harvest of 2014 and 2015; “Dry_tons_per_acre”, dry tons per acre;
 730 “Dry_Weight”, dry weight of biomass; “Lignin”, lingnin content; “NDF”, average neutral detergent fiber content of 2014 and
 731 2015; “NFC”, average non-fibrous carbohydrates content of 2014 and 2015; “Pericarp_pigmentation”, pericarp pigmentation;
 732 “Photoperiod_sensitivity”, photoperiod sensitivity; “Sorghum_type”, sorghum variety type: sweet, grain and cellulosic;
 733 “Stalk_height”, stalk height; “Total_fresh_weight”, total fresh weight; “WSC”, average water-soluble carbohydrates content of
 734 2014 and 2015. Data labels on the top of each tripartite column set indicate the percentage of SVs (the value before the colon)
 735 and SNPs (the value behind the colon) detected in SNP+SV GWAS that were also detected in SV GWAS or SNP GWAS. NA
 736 means that there was no signal detected in SNP+SV GWAS. The number of identified signals in SNP GWAS was always the
 737 lowest compared with other datasets for all phenotypes. The detected SNPs signals in SNP+SV GWAS were mostly overlapped
 738 in the results of SNP GWAS. However, SVs detected in SV and SNP-SV GWAS were far from identity. This indicates that
 739 association analysis based on all three of SNP, SV and integrated SNP+SV datasets is necessary to dissect genetic mechanisms
 740 thoroughly.

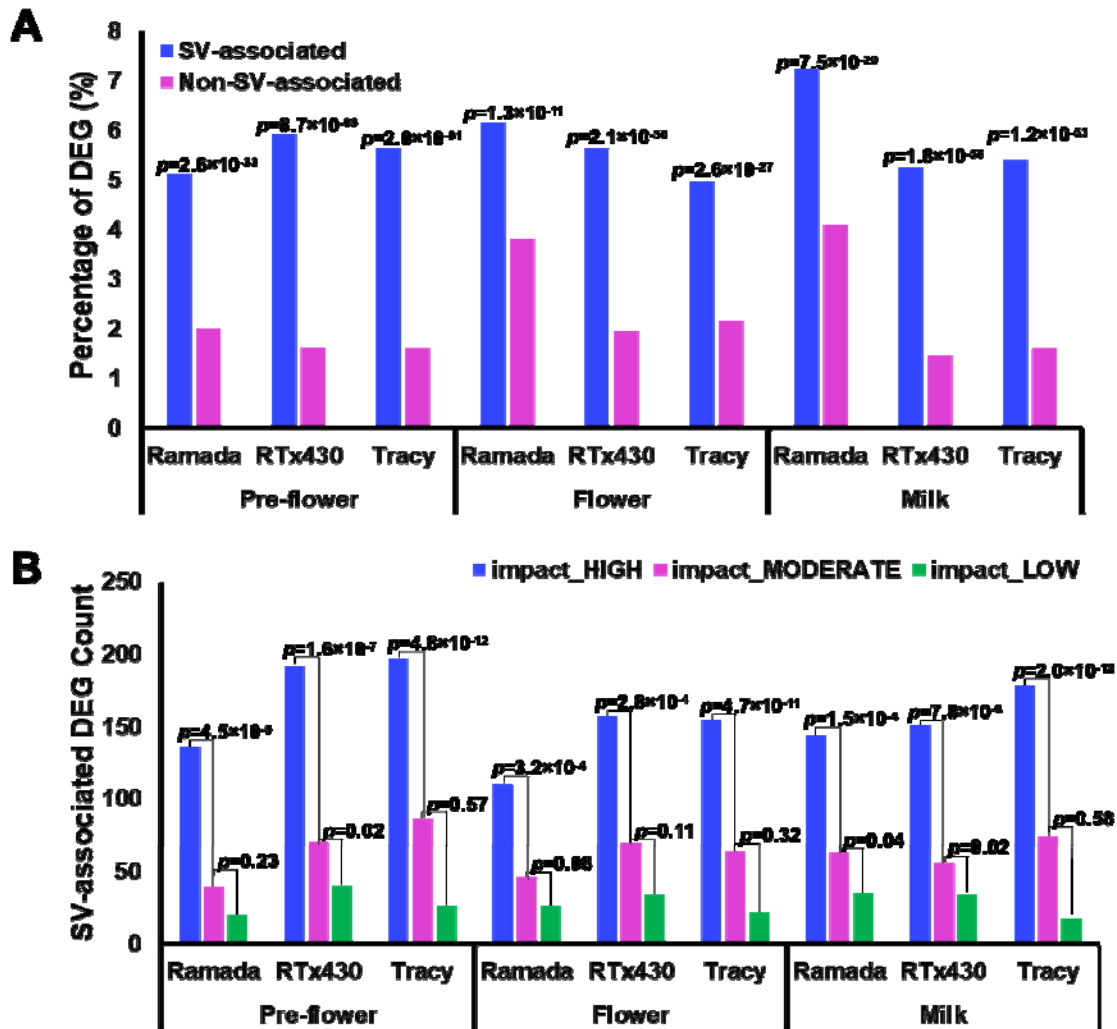
741

742

743

744

745



746

747 **Figure 8** Structural variants (SVs) have a widespread impact on gene expression. A SVs have an impact on gene expression in
 748 sorghum leaf across all developmental stages. The differentially expressed gene (DEG) analysis was performed by comparison of
 749 expression profiles in RTx430, Tracy and Ramada with the expression profile in Tx623 (as control) in leaf tissue at three
 750 developmental stages. Blue and pink bars represent the percentages of SV-associated and non-SV-associated DEGs respectively.
 751 The *p* values on the top of SV-associated DEG bars, which were adjusted using Bonferroni correction, indicate the
 752 hypergeometric testing results for enrichment of DEGs in SV-associated genes. DEGs were significantly enriched in SV-
 753 associated genes, with SV-associated DEGs increased 1.1%~4.3% compared with non-SV-associated DEGs in different sorghum
 754 lines. Only the results in leaf tissue were showed here. Similar results were also observed in stem tissue (see Supplemental Fig
 755 S11A-B). **B** SV-associated DEG count changed according to different impact predictions. Different classes of variant effects
 756 were predicted by SnpEff (v5.0)⁵⁵. The vertical axis showed the SV-associated DEG count. Blue, pink and green bars represent
 757 the DEG counts associated by high impact SVs (impact_HIGH), moderate impact SVs (impact_MODERATE) and low impact
 758 SVs (impact_LOW) respectively in leaf tissue of different sorghum lines in three developmental stages (pre-flower, flower and
 759 milk). The *p* values show the significance levels between groups (see Methods). Differential DEG counts between
 760 “impact_HIGH” and “impact_MODERATE” were all statistically significant. Significant level of DEG counts between
 761 “impact_MODERATE” and “impact_LOW” varied depending on lines and stages. In general, higher impact SVs associated
 762 more DEGs.

763

764

765 **References**

- 766 Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: an approach to discover, genotype, and
767 characterize typical and atypical CNVs from family and population genome sequencing. *Genome*
768 *Res* **21**: 974-984.
- 769 Alaei-Mahabadi B, Bhadury J, Karlsson JW, Nilsson JA, Larsson E. 2016. Global analysis of somatic
770 structural genomic alterations and their impact on gene expression in diverse human cancers.
771 *Proc Natl Acad Sci U S A* **113**: 13768-13773.
- 772 Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev*
773 *Genet* **12**: 363-376.
- 774 Allaby RG, Fuller DQ, Brown TA. 2008. The genetic expectations of a protracted model for the origins of
775 domesticated crops. *Proceedings of the National Academy of Sciences* **105**: 13982-13986.
- 776 Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren
777 D et al. 2020. Major Impacts of Widespread Structural Variation on Gene Expression and Crop
778 Improvement in Tomato. *Cell* **182**: 145-161 e123.
- 779 Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML,
780 Nelson BJ, Shah A, Dutcher SK et al. 2019. Characterizing the Major Structural Variant Alleles
781 of the Human Genome. *Cell* **176**: 663-675 e619.
- 782 Ayalew H, Peiris S, Chiluwal A, Kumar R, Tiwari M, Ostmeyer T, Bean S, Jagadish SVK. 2022. Stable
783 sorghum grain quality QTL were identified using SC35 x RTx430 mapping population. *Plant*
784 *Genome* **15**: e20227.
- 785 Boatwright JL, Sapkota S, Jin H, Schnable JC, Brenton Z, Boyles R, Kresovich S. 2022. Sorghum
786 Association Panel whole-genome sequencing establishes cornerstone resource for dissecting
787 genomic diversity. *Plant J* **111**: 888-904.
- 788 Brenton ZW, Cooper EA, Myers MT, Boyles RE, Shakoor N, Zielinski KJ, Rauh BL, Bridges WC,
789 Morris GP, Kresovich S. 2016. A Genomic Resource for the Development, Improvement, and
790 Exploitation of Sorghum for Bioenergy. *Genetics* **204**: 21-33.
- 791 Brenton ZW, Juengst BT, Cooper EA, Myers MT, Jordan KE, Dale SM, Glaubitz JC, Wang X, Boyles
792 RE, Connolly EL et al. 2020. Species-Specific Duplication Event Associated with Elevated
793 Levels of Nonstructural Carbohydrates in Sorghum bicolor. *G3 (Bethesda)* **10**: 1511-1520.
- 794 Browning BL, Zhou Y, Browning SR. 2018. A One-Penny Imputed Genome from Next-Generation
795 Reference Panels. *Am J Hum Genet* **103**: 338-348.
- 796 Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*
797 **34**: i884-i890.
- 798 Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Kallberg M, Cox AJ, Kruglyak S,
799 Saunders CT. 2016. Manta: rapid detection of structural variants and indels for germline and
800 cancer sequencing applications. *Bioinformatics* **32**: 1220-1222.
- 801 Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using
802 phased assembly graphs with hifiasm. *Nat Methods* **18**: 170-175.
- 803 Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L, Consortium GT
804 et al. 2017. The impact of structural variation on human gene expression. *Nat Genet* **49**: 692-699.
- 805 Chiluwal A, Perumal R, Poudel HP, Muleta K, Ostmeyer T, Fedenia L, Pokharel M, Bean SR, Sebela D,
806 Bheemanahalli R et al. 2022. Genetic control of source-sink relationships in grain sorghum.
807 *Planta* **255**: 40.
- 808 Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM, 2012. A
809 program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff:
810 SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *fly*, 6(2), pp.80-92.
- 811 Climente-González H, Azencott CA, Kaski S, Yamada M. 2019. Block HSIC Lasso: model-free
812 biomarker detection for ultra-high dimensional data. *Bioinformatics*, 35(14), pp.i427-i435.
- 813 Cooper EA, Brenton ZW, Flinn BS, Jenkins J, Shu S, Flowers D, Luo F, Wang Y, Xia P, Barry K et al.
814 2019. A new reference genome for Sorghum bicolor reveals high levels of sequence similarity

815 between sweet and grain genotypes: implications for the genetics of sugar metabolism. *BMC*
816 *Genomics* **20**.

817 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth
818 GT, Sherry ST et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156-2158.

819 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G,
820 Rivas MA, Hanna M et al. 2011. A framework for variation discovery and genotyping using next-
821 generation DNA sequencing data. *Nat Genet* **43**: 491-498.

822 Deschamps S, Zhang Y, Llaca V, Ye L, Sanyal A, King M, May G, Lin H. 2018. A chromosome-scale
823 assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nat Commun*
824 **9**: 4844.

825 Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. *Nat Rev Genet* **7**: 85-
826 97.

827 Figueiredo LF, Sine B, Chantreau J, Mestres C, Fliedel G, Rami JF, Glaszmann JC, Deu M, Courtois B.
828 2010. Variability of grain quality in sorghum: association with polymorphism in Sh2, Bt2, SssI,
829 Ae1, Wx and O2. *Theor Appl Genet* **121**: 1171-1185.

830 Fliege CE, Ward RA, Vogel P, Nguyen H, Quach T, Guo M, Viana JPG, Dos Santos LB, Specht JE,
831 Clemente TE et al. 2022. Fine mapping and cloning of the major seed protein quantitative trait
832 loci on soybean Chromosome 20. *Plant J* **110**: 114-128.

833 Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-
834 Smith C, Hurles ME et al. 2006. Copy number variation: new insights in genome diversity.
835 *Genome Res* **16**: 949-961.

836 Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome
837 assemblies. *Bioinformatics* **29**: 1072-1075.

838 Hamblin MT, Mitchell SE, White GM, Gallego J, Kukatla R, Wing RA, Paterson AH, Kresovich S. 2004.
839 Comparative Population Genetics of the Panicoid Grasses: Sequence Polymorphism, Linkage
840 Disequilibrium and Selection in a Diverse Sample of Sorghum bicolor. *Genetics* **167**: 471-483.

841 Hu Z, Olatoye MO, Marla S, Morris GP. 2019. An Integrated Genotyping-by-Sequencing Polymorphism
842 Map for Over 10,000 Sorghum Genotypes. *Plant Genome* **12**.

843 Ibraheem F, Gaffoor I, Chopra S. 2010. Flavonoid phytoalexin-dependent resistance to anthracnose leaf
844 blight requires a functional yellow seed1 in Sorghum bicolor. *Genetics* **184**: 915-926.

845 Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bahler J, Sedlazeck FJ.
846 2017. Transient structural variations have strong effects on quantitative traits and reproductive
847 isolation in fission yeast. *Nat Commun* **8**: 14061.

848 Kamvar ZN, Tabima JF, Grunwald NJ. 2014. Poppr: an R package for genetic analysis of populations
849 with clonal, partially clonal, and/or sexual reproduction. *PeerJ* **2**: e281.

850 Kendig KI, Baheti S, Bockol MA, Drucker TM, Hart SN, Heldenbrand JR, Hernaez M, Hudson ME,
851 Kalmbach MT, Klee EW et al. 2019. Sentieon DNaseq Variant Calling Workflow Demonstrates
852 Strong Computational Performance and Accuracy. *Front Genet* **10**: 736.

853 Kruglyak L. 2008. The road to genome-wide association studies. *Nature Reviews Genetics* **9**: 314-318.

854 Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos:
855 an information aesthetic for comparative genomics. *Genome Res* **19**: 1639-1645.

856 Lee T-H, Guo H, Wang X, Kim C, Paterson AH. 2014. SNPhylo: a pipeline to construct a phylogenetic
857 tree from huge SNP data. *BMC Genomics* **15**: 162.

858 Li C, Xiang X, Huang Y, Zhou Y, An D, Dong J, Zhao C, Liu H, Li Y, Wang Q et al. 2020. Long-read
859 sequencing reveals genomic structural variations that underlie creation of quality protein maize.
860 *Nat Commun* **11**: 17.

861 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome
862 Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools.
863 *Bioinformatics* **25**: 2078-2079.

864 Li Y, Varala K, Moose SP, Hudson ME. 2012. The inheritance pattern of 24 nt siRNA clusters in
865 *Arabidopsis* hybrids is influenced by proximity to transposable elements. *PLoS One* **7**: e47043.

866 Lipka AE, Kandianis CB, Hudson ME, Yu J, Drnevich J, Bradbury PJ, Gore MA. 2015. From association
867 to prediction: statistical methods for the dissection and selection of complex traits in plants. *Curr*
868 *Opin Plant Biol* **24**: 110-118.

869 Liu H, Liu H, Zhou L, Zhang Z, Zhang X, Wang M, Li H, Lin Z. 2015. Parallel Domestication of the
870 Heading Date 1 Gene in Cereals. *Mol Biol Evol* **32**: 2726-2737.

871 Lobell DB, Burke MB, Tebaldi C, Mastrandrea MD, Falcon WP, Naylor RL. 2008. Prioritizing Climate
872 Change Adaptation Needs for Food Security in 2030. *Science* **319**: 607–610.

873 Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq
874 data with DESeq2. *Genome Biol* **15**: 550.

875 Lu S, Dong L, Fang C, Liu S, Kong L, Cheng Q, Chen L, Su T, Nan H, Zhang D et al. 2020. Stepwise
876 selection on homeologous PRR genes controlling flowering and maturity during soybean
877 domestication. *Nat Genet* **52**: 428-436.

878 Mace ES, Tai S, Gilding EK, Li Y, Prentis PJ, Bian L, Campbell BC, Hu W, Innes DJ, Han X et al. 2013.
879 Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop
880 sorghum. *Nat Commun* **4**: 2320.

881 Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: A fast and
882 versatile genome alignment system. *PLoS Comput Biol* **14**: e1005944.

883 Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham
884 RK et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature*
885 **470**: 59-65.

886 Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ,
887 Acharya CB, Mitchell SE et al. 2013a. Population genomic and genome-wide association studies
888 of agroclimatic traits in sorghum. *Proc Natl Acad Sci U S A* **110**: 453-458.

889 Morris GP, Rhodes DH, Brenton Z, Ramu P, Thayil VM, Deshpande S, Hash CT, Acharya C, Mitchell
890 SE, Buckler ES et al. 2013b. Dissecting genome-wide association signals for loss-of-function
891 phenotypes in sorghum flavonoid pigmentation traits. *G3 (Bethesda)* **3**: 2085-2094.

892 Nattestad M, Schatz MC. 2016. Assemblytics: a web analytics tool for the detection of variants from an
893 assembly. *Bioinformatics* **32**: 3021-3023.

894 O'Donnell S, Fischer G. 2020. MUM&Co: accurate detection of all SV types through whole-genome
895 alignment. *Bioinformatics* **36**: 3242-3243.

896 Pawel S and James RL, 2010. Structural variation in the human genome and its role in disease. Annual
897 review of medicine, 61, pp.437-455.

898 Pelèse-Siebenbourg F, Caelles C, Kader J-C, Delseny M, Puigdomènech P. 1994. A pair of genes coding
899 for lipid-transfer proteins in *Sorghum vulgare*. *Gene* **148**: 305-308.

900 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI,
901 Daly MJ et al. 2007. PLINK: a tool set for whole-genome association and population-based
902 linkage analyses. *Am J Hum Genet* **81**: 559-575.

903 R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for
904 Statistical Computing.

905 Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO. 2012. DELLY: structural variant
906 discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**: i333-i339.

907 Rhodes DH, Hoffmann L, Jr., Rooney WL, Ramu P, Morris GP, Kresovich S. 2014. Genome-wide
908 association study of grain polyphenol concentrations in global sorghum [*Sorghum bicolor* (L.)
909 Moench] germplasm. *J Agric Food Chem* **62**: 10916-10927.

910 Sandesh K, Ujwal P. 2021. Trends and perspectives of liquid biofuel – Process and industrial viability.
911 *Energy Conversion and Management: X* **10**.

912 Scherer, S.W., Lee, C., Birney, E., Altshuler, D.M., Eichler, E.E., Carter, N.P., Hurles, M.E. and Feuk, L.,
913 2007. Challenges and standards in integrating surveys of structural variation. *Nature genetics*,
914 39(Suppl 7), pp.S7-S15.

915 Shen W, Le S, Li Y, Hu F. 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File
916 Manipulation. *PLoS One* **11**: e0163962.

917 Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing
918 genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**:
919 3210-3212.

920 Simms D, Cizdziel PE, Chomczynski P. TRIzol: A new reagent for optimal single-step isolation of RNA.
921 Focus 15, 532-535 (1993).

922 Song Q, Zhang T, Stelly DM, Chen ZJ. 2017. Epigenomic and functional analyses reveal roles of
923 epialleles in the loss of photoperiod sensitivity during domestication of allotetraploid cottons.
924 *Genome Biol* **18**: 99.

925 Song R, Llaca V, Messing J. 2002. Mosaic organization of orthologous sequences in grass genomes.
926 *Genome Res* **12(10)**: 1549-1555.

927 Songsomboon K, Brenton Z, Heuser J, Kresovich S, Shakoore N, Mockler T, Cooper EA. 2021. Genomic
928 patterns of structural variation among diverse genotypes of *Sorghum bicolor* and a potential role
929 for deletions in local adaptation. *G3 (Bethesda)* **11**.

930 The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale
931 sequencing. *Nature* **467**: 1061-1073.

932 Urriola J, Rathore KS. 2015. Overexpression of a glutamine synthetase gene affects growth and
933 development in sorghum. *Transgenic Res* **24**: 397-407.

934 Villanueva RAM, Chen ZJ. 2019. ggplot2: elegant graphics for data analysis. Taylor & Francis.

935 Wang J, Zhang Z. 2021. GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and
936 Prediction. *Genomics Proteomics Bioinformatics* doi:10.1016/j.gpb.2021.08.005.

937 Yang N, Liu J, Gao Q, Gui S, Chen L, Yang L, Huang J, Deng T, Luo J, He L et al. 2019. Genome
938 assembly of a tropical maize inbred line provides insights into structural variation and crop
939 improvement. *Nat Genet* **51**: 1052-1059.

940 Zarate S, Carroll A, Mahmoud M, Krasheninina O, Jun G, Salerno WJ, Schatz MC, Boerwinkle E, Gibbs
941 RA, Sedlazeck FJ. 2020. Parliament2: Accurate structural variant calling at scale. *Gigascience* **9**.

942 Zhang LM, Leng CY, Luo H, Wu XY, Liu ZQ, Zhang YM, Zhang H, Xia Y, Shang L, Liu CM et al.
943 2018. Sweet Sorghum Originated through Selection of Dry, a Plant-Specific NAC Transcription
944 Factor Gene. *Plant Cell* **30**: 2286-2307.

945 Zhang Q, Prive F, Vilhjalmsson B, Speed D. 2021. Improved genetic prediction of complex traits from
946 individual-level data or summary statistics. *Nat Commun* **12**: 4192.

947 Zhang R, Jia G, Diao X. 2023a. geneHapR: an R package for gene haplotypic statistics and visualization.
948 *BMC Bioinformatics* **24**: 199.

949 Zhang Y, Hu Y, Wang Z, Lin X, Li Z, Ren Y, Zhao J. 2023b. The translocase of the inner mitochondrial
950 membrane 22-2 (AtTIM22-2) is required for mitochondrial membrane functions during
951 *Arabidopsis* seed development. *J Exp Bot* doi:10.1093/jxb/erad141.

952 Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas
953 JM et al. 2010. Mixed linear model approach adapted for genome-wide association studies. *Nat*
954 *Genet* **42**: 355-360.

955 Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. 2012. A high-performance computing
956 toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**: 3326-
957 3328.

958