Supplemental Material for:

Major impacts of widespread structural variation on sorghum

Zhihai Zhan, Joao Paulo Gomes Viana, Bosen Zhang, Kimberly K.O. Walden, Hans Müller Pau, Stephen P Moose, Geoff Morris, Chris Daum, Kerrie W Barry, Nadia Shakoor, Matthew E Hudson.

Supplemental Results

Simulation study of recall and precision in SV calling

To enhance the accuracy and sensitivity of SV detection, five inference software packages: Sentieon (v202010.01) (Kendig et al. 2019), DELLY (v0.8.1) (Rausch et al. 2012), Smoove (https://github.com/brentp/smoove), manta (v1.6.0) (Chen et al. 2016) and CNVnator (v0.3.3) (Abyzov et al. 2011), involving different SV detection strategies, were applied to the data. To achieve a balance between recall rate and precision in SV calling with diverse supported callers, we conducted a simulation study which aimed to access recall and precision in SV calling across various thresholds, taking into account SVs supported by one to five callers. To retain the genomic characteristics of sorghum, we simulated the variations based on the Chromosome 1 and Chromosome 2 of the standard sorghum reference genome BTx623 (v3.1.1) using Mason (v2.0.9) software (Holtgrewe, 2010). SNP and small indel rates were respectively set to 0.0001 and 0.000001. The rates of all SV types, including DEL, INS, INV, DUP, INV and TRA, were set to 0.0000005. Consequently, there were 15,477 SNVs (SNPs + small Indels), 384 BNDs, 43 DELs, 73 DUPs, 43 INSs, and 77 INVs generated in the simulated variants. 30× NGS pair-reads (34 million Illumina short-reads (150 bp)) were simulated. The pipeline used for variant calling in the BAP was then utilized for the simulated reads. Because Mason can only simulate multiple BNDs to represent the breakends of TRAs, which was manifested through the fact that the number of BNDs (384) was far higher than the other SVs even though we set the same rate for all SVs, we compared all breakpoints between the simulated SVs and the identified SVs using our ensemble pipeline to calculate the true positive, false positive, and false negative rates in different fusion thresholds using a home-made script. Generally, the precision rate increases while the recall rate decreases with the increment of the supported callers (Supplemental Fig S1). The recall and precision trendlines cross between the “supported by at least 1 caller” and “supported by at least 2 callers” points, which indicates the balance point of the recall and precision.

In current developed NGS-based tools, various algorithms differ in their strength and weakness in identifications of different SV types as mentioned above. Support by more callers will lead to more accuracy in SV calling, while also increasing type II error. Based on the result of the balancing study of recall and precision, we consider “supported by at least 2 callers” is a reasonable threshold for the SV fusion workflow in our study.

Structural variant calling and validation of fusion workflow

An average of 90.26% DEL/INS (PI 329545, 91.47%; PI 337680, 96.10%; PI 651495, 88.78%; Rio, 82.57%; RTx430, 92.37%) and 85.91% DUP (PI 329545, 93.74%; PI 337680, 97.22%; PI 651495, 86.34%; Rio, 59.16%; RTx430, 93.09%) were identified using our fusion workflow at overlapping genomic positions to the SVs identified using assembly comparison, after filtering out very large calls (SVs larger than 10 Mb) from assembly comparison (see Supplemental Fig S3). Due to the limitation of the SV detection algorithms based on next-generation sequencing reads, the sequence details of INVs and TRAs cannot be inferred accurately based on their two breakpoints (Pawel Stankiewicz and James R. Lupski, 2010). To evaluate the accuracy of TRAs and INVs, we developed a custom python script to capture the consistency of breakpoints of our identified TRAs and INVs with the spanning of the presence/absence type variants identified by assembly comparisons of the three chromosome-scale de novo assemblies and the two public whole genome sequence assemblies included in the BAP. An average of 76.64% INVs (PI 329545, 86.45%; PI 337680, 92.78%; PI 651495, 80.70%; Rio, 37.12%; RTx430, 86.15%) and 91.51% TRAs (PI 329545, 94.67%; PI 337680, 99.26%; PI 651495, 89.36%; Rio, 81.53%; RTx430, 92.74%) were consistent with the presence/absence variant pairs. Even though only 59.16% DUPs and 37.12% INVs were covered in Rio, the concordance ratio of DEL/INS and TRA in Rio, and SVs in other accessions were consistently more than 80%. It is important to note that the deviation between the SV datasets could attribute to the high-complexity genome, potential errors in genome assembly, the whole genome alignment algorithm, or any other intractable factors. These results manifests that the footprints of most our identified TRAs and INVs were traceable from the corresponding presence/absence variants alleles located in two breakpoints of TRAs/INVs.

Songsomboon et al. identified 22,359 deletions and 2,009 duplications in 347 diverse sorghum genotypes included in the BAP and characterized their genomic patterns in diversity and potential role in local adaptation (Songsomboon et al. 2021). In order to further verify the reliability of our SV calling workflow, we also compared the published deletions and duplications with the DELs and DUPs identified in our study. Over 90% and 91% of their published deletions and duplications were present in our DEL and DUP datasets, respectively. These high concordances with reference-based variant calling demonstrate that our fusion pipeline accurately identifies genome-wide SVs.

Structural variant data allows detection of new GWAS associations

GWAS based on SV for "Pericarp_pigmentation" trait found three significant association signals for seed pericarp pigmentation, including an SV underlying the *Y1* gene (*Sobic.001G397900*) as expected (Figure 6A), while SNP analysis did not detect association at this locus (Figure 6B). The SV (1.5 kb downstream of *Y1*) underlying the *Y1* locus was called as a translocation from 68,368,772 bp on Chromosome 1 to 67,961,157 bp on Chromosome 4. After checking the adjacent SVs, we found there were three additional TRAs called from 68,368,773 bp on Chromosome 1 to 67,950,235 bp on Chromosome 4 by different algorithms. We also identified two TRAs called in the opposite direction: one from at 67,950,235 bp on Chromosome 4 to 68,368,773 bp on Chromosome 1, the other from 67,961,154 bp on Chromosome 4 to 68,368,775 bp on Chromosome 1. These reciprocal breakpoints called by software at this locus are most likely transposon hotspots spanning at least 10.9 kb (67,950,235 – 67,961,157 bp on Chromosome 4) moved between Chromosome 4 and Chromosome 1 completely or partially (Figure 6D). Importantly, a significant pericarp-pigmentation associated TRA with the start breakpoint at 67,961,154 bp on Chromosome 4 was also detected, which is within the putative transposon hotspot above. These findings are likely to indicate novel causative alleles of the seed pericarp pigmentation trait.

In order to validate the putative transposon hotspots at the detected loci between Chromosome 1 and Chromosome 4, we examined the sequence composition of the putative transposon regions in five sorghum lines with available genome assemblies (RTx430, Rio, PI 329545, PI 337680 and PI 651495). Based on our called SV set, the breakpoints of the TRA at 68,368,772 bp on Chromosome 1 and 67,961,154 bp on Chromosome 4 were only detected in RTx430, PI 329545, PI 337680 and PI 651495. Therefore, we scanned the potential transposon insertions covering the breakpoint at 68,368,772 bp on Chromosome 1 and the 10.9 kb region from 67,950,235 to 67,961,157 bp on Chromosome 4 in the five genomes compared with BTx623 using a whole genome alignment approach. As expected, there were not any potential insertions encompassing the breakpoint and the 10.9 kb region in Rio. In RTx430, there was a 37.8 kb genomic fragment detected covering the breakpoint at 68,368,772 bp on Chromosome 1 (Supplemental Table S8). Repeat annotation indicated that the 37.8 kb fragment contained a transposon cluster spanning about 37.7 kb (Supplemental Table S9). A 19.8 kb genomic fragment covering the breakpoint at 68,368,772 bp on Chromosome 1 was identified in PI 329545, which overlapped with a 19.7 kb transposon rich area (Supplemental Table S8, S9). In PI 337680, we also found a 19.8 kb insertion covering the breakpoint at 68,368,772 bp on Chromosome 1, which overlapped with an annotated transposon enrichment region spanning 19.7 kb (Supplemental Table S8, S9). Meanwhile, PI 337680 also contains an 8.3 kb transposon-rich insertion in the putative transposon hotspot region on Chromosome 4 (Supplemental Table S8, S9). In PI 651495, a 16.6 kb transposon enrichment insertion covering the breakpoint at 68,368,772 bp on Chromosome 1 was detected (Supplemental Table S8). These results suggested that the regions underlying the reciprocal breakpoints between Chromosome 4 and Chromosome 1 are transposon hotspots.

In order to validate the significance of the TRA allele underlying the *Y1* locus on Chromosome 1, we performed haplotype analyses. There were three main haplotypes (each haplotype contained at least five sorghum lines) identified, and H002 (with the TRA allele haplotype) showed significant phenotypic difference with H001 (without the TRA allele haplotype, $p = 6.4^{-8}$) and H003 (without the TRA allele haplotype, $p = 0.022$) (Supplemental Fig S9A, B). Allele effect estimation analysis indicated that the TRA allele has the second highest effect of all included alleles, and the highest -$\log_{10}$ (*p*-value) for “Pericarp_pigmentation” (Supplemental Fig

S9C). Haplotype network analysis indicated that the translocation event likely predated the variety type differentiation (Supplemental Fig S9D). These data illustrate the power boost and increased interpretability, details and complexity brought forth by inclusion of SVs in GWAS analysis.

Another substantial SV association signal for seed pericarp pigmentation was detected on Chromosome 8. The polymorphism associated at this locus is a 2.6 kb DEL/INS located 3.2 kb upstream of *TIM22-2* (*Sobic.008G111800*), a mitochondrial import inner membrane translocase and a homolog of a protein involved in seed development in *Arabidopsis* (Zhang et al. 2023b). To validate the significance of the DEL variation, we again conducted haplotype effect analyses. Seven primary haplotypes (each haplotype contained at least five sorghum lines) were identified, and only one haplotype, H002, did not contain the DEL allele (Supplemental Fig S10A). Phenotypic comparison of "Pericarp_pigmentation" between different haplotypes (each haplotype contained at least five sorghum lines with available phenotypic data for statistically comparisons) showed that H002 displayed a significant phenotypic difference with all other haplotypes (Supplemental Fig S10B). Furthermore, this DEL showed the highest -$\log_{10}$(*p*-value) and third highest allelic effect (Supplemental Fig S10C). Haplotype network analysis indicated that this deletion may have occurred before the variety type differentiation (Supplemental Fig S10D). Importantly, the significant locus was only found when SV GWAS alone was conducted; there was no signal exceeding the Bonferroni corrected threshold in our GWAS for the "pericarp_pigmentation" trait based on either the SNP and SNP+SV analyses, even though the same SVs on Chromosome 1 and Chromosome 4 detected in SV-based GWAS were still within the top 10 loci associated with "pericarp pigmentation" in GWAS based on SNP+SV (Figure 6B, 6C). The addition of SNPs to SV GWAS analysis gives only slightly increased heritability estimation of the “pericarp_pigmentation” trait, from 24.1% (SV only) to 26.2% (SNP+SV). This small increase from the addition of SNPs is likely insufficient to overcome the statistical power penalty in GWAS of the increased multiple testing correction. The Bonferroni corrected threshold ($\alpha = 0.05$) is based on 589,579 polymorphisms for SV analysis, but 7,735,403 for SNP+SV analysis. Our GWAS results for seed pericarp pigmentation based on SVs thus not only found a significant SV association for the well-studied *Y1* locus, which was not detected in SNP GWAS, but also identified a potential translocation involved in the genesis of this locus and a compelling new candidate gene for the control of seed pericarp pigmentation.

SVs detected within the BAP are also important determinants of gene expression in other lines

Knowing that SVs affect gene expression within the BAP, we investigated whether the differentially expressed genes between cellulosic and sweet sorghum lines beyond the BAP could be identified by SVs identified between our representative cellulosic and sweet sorghum lines. To explore the broad influence, we additionally performed RNA-seq on 4 typical cellulosic (TAM08001, TAM17500, TAM17600, TAM17800) and 6 typical sweet sorghum lines (Brawley, GBR, M81E, R9188, Topper, Tracy) which are not included in the 363 lines except for Brawley, Topper and Tracy. For each development stage and tissue, DEGs were identified between the cellulosic group (TAM08001, TAM17500, TAM17600, TAM17800 in same tissue and stage) and sweet group (Brawley, GBR, M81E, R9188, Topper, Tracy in same tissue and stage) and relative levels of SV-associated DEGs and non-SV-associated DEGs evaluated. Only CNV type variations were considered and the CNVs that were present in both cellulosic group and sweet group were excluded. The proportions of SV-associated DEGs are still much higher than non-SV-associated DEGs in all stages and tissues (Supplemental Fig S13). The SVs identified between representative cellulosic or sweet sorghums in BAP also showed an effect on genes expression in a cellulosic-sweet sorghum comparison beyond BAP (Supplemental Fig S14). These findings suggest that our identified SVs between our curated cellulosic and sweet sorghum lines are also important in other cellulosic-sweet comparisons.

To further validate the feasibility of prediction from our identified SVs to differential gene expression between cellulosic and sweet sorghum lines beyond the BAP, we formulated the prediction model based on Block Hilbert Schmidt Independence Criterion Lasso (Block HSIC Lasso) method (Climente-González et al. 2019). CNV type variations were used as independent variable in prediction models. Genes with adjusted p value < 0.05 and log2FoldChange > 2 were defined as DEGs. Area under the receiver operating characteristics (AUROC) score was used to assess the fitness of the prediction models. The prediction model showed promising precision, recall, F1, accuracy and AUROC scores among across all stages and tissues (Supplemental Fig S15, Supplemental Table S17). These results indicate that DEGs between other cellulosic-sweet comparisons can be predicted by the SVs identified from the BAP across all stages and tissues (Supplemental Fig S15, Supplemental Table S17).

Supplemental Methods

Re-sequencing dataset and phenotypes

The Illumina short-read sequence dataset and phenotypes of the sorghum lines used in this study were collected by the TERRA-REF project http://terraref.org (Brenton et al. 2016), which consists of 390 sorghum lines that represent a wide range of molecular and phenotypic diversity. Sequence data was available for 363 sorghum lines. Only 339 sorghum lines with population information were considered for population genetic analysis. Sorghum information for each line is included in Supplemental Table S1.

Variant calling

The raw FASTQ files were cleaned by fastp (version 0.20.0) software (Chen et al. 2018) using the default parameters. Cleaned reads were mapped to the latest sorghum BTx623 (v3.1.1) from phytozome (https://phytozome.jgi.doe.gov/). Single-nucleotide polymorphisms (SNPs) were called using the Sentieon (version 202010.01) (Kendig et al. 2019) DNA-seq pipeline. There were 38,325,772 were detected in total across all 10 chromosomes initially. Only bi-allelic SNPs were kept following quality filtering criteria: QD < 2.0, MQ < 40.0, MQRankSum < -12.5, ReadPosRankSum < -8.0, FS > 60.0, SOR > 3.0 by GATK (v4.1.4.0) (DePristo et al. 2011). Genotype value filtering and allele filtering were executed by VCFtools(v0.1.15) (Danecek et al. 2011) using --minQ 30 --minDP 30 and --maf 0.05 –max-missing 0.95. The imputation process was conducted by Beagle (v5.1) (Browning et al. 2018). The SNPs data set was pruned by plink (v1.90) (Purcell et al. 2007) using --indep-pairwise 20 5 0.8, which means the windows size is 20kb, which is based on the established linkage disequilibrium range for sorghum published in the previous study (Hamblin et al. 2004; Mace et al. 2013), shifting step size was 5 SNPs, one of a pair of SNPs will be removed if the LD is greater than 0.8 (by r^2). A final total of 7,162,000 SNP was kept for the subsequent analyses. Five independent tools based on different algorithms were used to call structural variations (SVs): Sention (version 202010.01) DNAscope algorithm (Kendig et al. 2019), DELLY (v0.0.1, -q 20) (Rausch et al. 2012), Smoove (v0.2.7), manta (v1.6.0) (Chen et al. 2016) and CNVnator (v0.3.3) (Abyzov et al. 2011). SVs per individual from each caller were fused by SURVIVOR (Jeffares et al. 2017) considering consistency of SVs type and SVs strands. The SVs called from the different platforms with breakpoints within 1kb were fused. Fusion SVs datasets per individual were genotyped maximizing the yield of merged variants with a purpose-written python script, which avoids the bias genotyped by only one caller. Fusion SVs from each individual were merged by Jasmine (version 1.1.4) (https://github.com/mkirsche/Jasmine). SVs with consistent type and strands, and start positions, end positions within 1kb were merged into a single SV call. Population SVs raw dataset was created by merging SVs from 363 individuals. Only SVs meeting the SV length within 30bp-1Mb and MAF > 0.001 were kept. To further refine the SVs, we also removed the ungenotyped SVs, SVs from BTx623 and all their adjacent SVs within 2.5Kb extended flank region. There were 622,236 SVs kept on 10 chromosomes for the downstream analyses. Circle plot was drawn by circus (Krzywinski et al. 2009).

de novo assembly

Leaves from the seedlings of sorghum in greenhouse were sampled. At least 10g of leaf tissue for each sorghum accession was sent to Roy J. Carver Biotechnology Center at the University of Illinois at Urbana-Champaign. Raw HIFI sequence data in BAM format was generated by PacBio Sequel IIe platform. Generated sequencing BAM files were converted to FASTQ files by SAMtools (Li et al. 2009). Reads less than 1 kb were identified and filtered by SeqKit tools (Shen et al. 2016). Genome *de novo* assembly were performed by hifiasm (Cheng et al. 2021). Genome assembly quality was evaluated by quast (Gurevich et al. 2013), and BUSCO (Simao et al. 2015).

Evaluation of SV calling

Comparison between the Sorghum reference genome (BTx623, Sorghum bicolor V3.1.1, https://phytozome-next.jgi.doe.gov/) and two other assembled sorghum genomes, Rio (Sorghum bicolor Rio v2.1) and RTx430 (Sorghum bicolor RTx430 v2.1) downloaded from Phytozome (https://phytozome-next.jgi.doe.gov/) was used to evaluate the SV calling. MUM&Co(v3.7) (O'Donnell and Fischer 2020) was used to detect the SVs from Rio-BTx623 and RTx430-BTx623 through whole genome alignment information provided by MUMmer (v4) (Marcais et al. 2018). The SVs called from the MUM&Co were then compared to SVs from the fusion pipeline (See Variant calling in Methods) to calculate the overlapped number of SVs using the custom-written python script *cnv_overlapped_calculation.py*.

Mobile elements annotation

BTx623 Repeatmask annotation file Sbicolor_454_v3.1.1.repeatmasked_assembly_v3.0.1.gff3.gz was downloaded from https://phytozome-next.jgi.doe.gov/. For rearrangement-type variations, both the start and end positions were considered for calculation of the overlap between the SV breakpoints and annotated mobile elements boundaries, by using a custom python script. For CNV type variations, we only considered whether the start position was located within the boundaries of annotated mobile elements.

Heritability estimation

LDAK (v5.1) (Zhang et al. 2021) was used to estimate the trait heritability explained by the SNP and SV polymorphisms. SNP data was pruned by plink 1.9 before LDAK was applied. First, variations were weighted by chromosomes by using --cut-weights; step 2, then a weighted kinship matrix generated by using --calc-kins-direct; finally, variance components were estimated by using --reml quant for quantitative traits and --reml binary for the binary trait. Plots were drawn in R 4.2.2 (R Core Team, 2022) by using the *ggplot2* package (Villanueva and Chen 2019).

Population genetics analysis

Principal component analysis was performed using the R function *prcomp()* (R Core Team, 2022). A minimum spanning tree was created using the R package *Poppr* (Kamvar et al. 2014). SNPhylo (Lee et al. 2014) was used to create maximum likelihood phylogenetic trees using following parameters: -l (LD threshold) 0.1, -m (MAF threshold): 0.1 for SNP dataset, 0 for SV dataset, -M (Missing rate): 0.1 for SNP dataset, 1 for SV dataset, -B (The number of bootstrap samples): 100, -b (Perform (non-parametric) bootstrap analysis and generate a tree), -A (Perform multiple alignment by MUSCLE), -r (Skip the step removing low quality data): only for SV dataset, the remaining parameters were default, after converting VCF to GDS format using R package *SNPRelate* (Zheng et al. 2012). SVs were converted to present-absent binary representation before conducting PCA. F_{ST} was calculated by using VCFtools (v0.1.16) (Danecek et al. 2011).

GWAS

GWAS was performed by GAPIT3 using the compressed mixed linear model (CMLM) model (Zhang et al. 2010; Wang and Zhang 2021). The first three principal components and a group kinship matrix calculated from clustered individuals were used for population structure correction. The gff3 file Sbicolor_454_v3.1.1.gene.gff3 was downloaded from phytozome (https://phytozome-next.jgi.doe.gov/).

Haplotype analyses

For the TRA underlying the *Y1* on Chromosome 1, the variations including SNP and SVs surrounding the breakpoint at 68,368,772 from 68,360,756 to 68,369,370 were used to identify the haplotype. For the 2.6 kb DEL on Chromosome 8, the variations including SNP and SVs surrounding the DEL from 51,753,931 to 51,758,539 were used to identify the haplotype. The R package *geneHapR* (Zhang et al. 2023a) was used to perform analyses.

RNA-seq analysis

Tissues samples for RNA were collected from plants grown in the field at the Energy Farm at the University of Illinois at Urbana-Champaign in 2018. Collected samples were ground in a pestle and mortar (for leaves) or a 6970 EFM Freezer/Mill (for stems, www.spexsampleprep.com/freezermill). RNA was extracted by using the Trizol and Chloroform methods (Simms et al. 1993) and precipitated using isopropanol. The primary quality control and concentration measurement of RNA was by using TURBO DNA-free™ Kit from Invitrogen, Qubit™ RNA BR Assay Kit from Thermo Fisher Scientific and Bioanalyzer from Agilent. Prepared RNA samples were sent to JGI (https://jgi.doe.gov/). There were three biological replicates. Samples with less than three successful biological replicates were discarded in the downstream analysis. RNA-seq data were analyzed by DESeq2 package (Love et al. 2014), and plot was drawn by ggplot2 (Villanueva and Chen 2019).

Analysis of association between SVs and gene expression

Analysis was performed in each line, each tissue, and each stage:
Step1, Calculate the significance level of genes expression difference against BTx623 samples in same tissue and developmental stage using DESeq2 package (Love et al. 2014). Create the gene expression matrixI, including 4 columns: "gene_id", "log2FoldChange", "p_value" and "Significance_level". "Significance_level" was assigned to "1" if meeting the criteria: |log2FoldChange| > 2 and adjusted "p_value" <0.05. Otherwise, the "Significance_level" was assigned to "0". "p_value" were adjusted using the Bonferroni correction. Only protein-coding genes on chromosomes were kept.
Step2, Predict the SV effects on sorghum genome using SnpEff (v5.0) (Cingolani et al. 2012) and BTx623 reference (Sorghum bicolor V3.1.1, https://phytozome-next.jgi.doe.gov/). Create the impact prediction matrixII based on the output files, including the following columns: "gene_id", "impact_HIGH", "impact_LOW", "impact_MODERATE", "impact_MODIFIER" and effect prediction details such as "3_prime_UTR_truncation", "disruptive_inframe_deletion" etc. which are pre-defined by the SnpEff software and varying depending on the effects of SV dataset. Only CNV type structural variations (DEL, DUP, INS) were taken into consideration in impact prediction because of the difficulty of associating other types of SV with specific genes. Only protein-coding genes on chromosomes were kept.

Step3, Combine the gene expression matrixI and the impact prediction matrixII according to the "gene_id" column, getting matrixIII. Genes that were not present in matrixII were filled with "0" in the effect annotation columns.

Step4, Create the discriminant matrixIV for all genes based on the combined matrixIII:

a) Add "Association" column, which was used to store the summary data to judge whether the DEG is associated with SV or not directly.

b) Create two sub-columns: "SV-associated" and "non-SV-associated" for "Association" column and all effect annotation columns as in matrixIII to store the data to judge whether the DEG is associated with SV or not in its corresponding predicted impact field.

c) Fill the matrixIV:
In "Association": "SV-associated" if sum of the number of all effect annotation fields is not equal to zero, which means the gene is associated by SV, and 1): if "Significance_level" is "1", the cell will be assigned "1", which means the DEG is associated by SV; 2): if "Significance_level" is "0", the cell will be assigned "0", which means the gene is associated by SV, but it is not a DEG. Otherwise, the cell will be assigned "NA", which means the gene can't be analyzed (lack of expression data or impact prediction data).
In "Association": "Non-SV-associated", if sum of the number of all effect annotation fields is equal to zero, which means the gene is not associated by SV, and 1): if "Significance_level" is "1", the cell will be assigned "1", which means the DEG is not associated by SV; 2): if "Significance_level" is "0", the cell will be assigned "0", which means the gene is not associated by SV, and it is not a DEG neither. Otherwise, the cell will be assigned "NA".
In predicted impacts columns, let's take "impact_HIGH" as an example. In "impact_HIGH": "SV-associated", if the cell in the corresponding column "impact_HIGH" in matrixIII is not "0", and 1): if "Significance_level" is "1", the cell will be assigned "1", which means the DEG is associated by SV and SV have a high impact on this DEG; 2): if "Significance_level" is "0", the cell will be assigned "0", which means the gene is associated by SV, and SV have a high impact on this gene, but it is not a DEG. Otherwise, the cell will be assigned "NA".
In "impact_HIGH": "Non-SV-associated", if the cell in the corresponding column in matrixIII is "0", and 1): if "Significance_level" is "1", the cell will be assigned "1", which means the DEG is not associated by SV in the "high impact" aspect; 2): if "Significance_level" is "0", the cell will be assigned "0", which means the gene is not associated by SV in the "high impact" aspect, and it is not a DEG neither. Otherwise, the cell will be assigned "NA".
The following effect annotation columns were filled as "impact_HIGH" column above.

Step5, Count the number of DEGs (labeled as "1") and "non-DEGs" (labeled as "0") in "SV-associated" and "Non-SV-associated" categories in "Association" and all effect prediction fields. Calculated the DEGs as percentage of all genes (for normalization of the total genes in "SV-associated" and "non-SV-associated" categories, ie, *DEGs as Percentage of all genes in SV-associated category = (DEG count in SV-associated category / ((DEG count + non-significant genes count) in SV-associated category). DEGs as Percentage of all genes in Non-SV-associated category = (DEG count in non-SV-associated category / ((DEG count + non-significant genes count) in Non-SV-associated category)*) in "SV-associated" and "Non-SV-associated" categories in "Association" and all effect prediction fields.

Step6, Hypergeometric testing for enrichment of DEGs in SV-associated genes. The number of SV-associated DEGs (SVD) follows the hypergeometric distribution:

$$\mathrm{SVD} \sim \mathrm{H}(\mathrm{n}, \mathrm{N}, \mathrm{M})$$

where n is the number of DEGs; N is the total number of the expressed genes; M is the number of SV-associated genes, with the following probability distribution:

$$f(k; N, M, n) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

where k is the number of SV-associated DEGs.

Statistical analysis is performed using *phyper()* function of *stats* package in R (R Core Team, 2022). “p-value” were under Bonferroni correction.

Step7, Statistical analysis of the significance level of the difference between SV-associated DEG counts (used as arrays composed by “0” and “1”) in “Impact_HIGH”, “Impact_MODERATE” and “Impact_LOW” fields using unpaired two-tailed T-test.

Supplemental Discussion

We were intrigued by the observation that the SVs are more abundant in gene-rich regions, and did not share the distribution of the SNPs. This result was indicated by the SV density calculated in 500 kb bins, showing that the SVs detected were primarily towards the gene-dense telomeres (main Figure 4). While this is an interesting observation, there are various methodological difficulties that most likely account for it. In particular, the repetitive nature of the centromeres and gene-poor repetitive regions surrounding them limits mapping quality of reads or the quality of alignments. Thus, the ability to call any variants in these regions is affected, but because mate pair mapping is needed for larger SV detection, this may be disproportionately affected. It is quite probable that one or more methodological reasons, rather than a genuine difference in polymorphism frequency, accounts for this observation; however, we do also expect to see more chromosome breakage towards telomeres, so the difference in distribution could be a genuine phenomenon. More research is needed to address this question.

Since the high abundance of SVs in the gene-rich regions appears to violate the principle of gene conservation, we investigated the proportion of SVs which directly affect exon regions. We found that, even though the density of called SVs is higher in gene-rich regions, only 0.2% of these SVs affected exons directly. Thus, this principle is still very much in force.

References

Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21**: 974-984.

Brenton ZW, Cooper EA, Myers MT, Boyles RE, Shakoor N, Zielinski KJ, Rauh BL, Bridges WC, Morris GP, Kresovich S. 2016. A Genomic Resource for the Development, Improvement, and Exploitation of Sorghum for Bioenergy. *Genetics* **204**: 21-33.

Browning BL, Zhou Y, Browning SR. 2018. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am J Hum Genet* **103**: 338-348.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**: i884-i890.

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Kallberg M, Cox AJ, Kruglyak S, Saunders CT. 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**: 1220-1222.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**: 170-175.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM, 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. fly, 6(2), pp.80-92.

Climente-González H, Azencott CA, Kaski S, Yamada M. 2019. Block HSIC Lasso: model-free biomarker detection for ultra-high dimensional data. Bioinformatics, 35(14), pp.i427-i435.

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

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

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

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

Holtgrewe M. 2010. Mason–a read simulator for second generation sequencing data. Technical Report FU Berlin.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Supplemental Table Legends

Supplemental Table S1 363 sorghum accessions

Supplemental Table S2 Summary statistics of the 3 assembled sorghum genomes

Supplemental Table S3 43 representative cellulosic accessions

Supplemental Table S4 33 representative sweet accessions

Supplemental Table S5 1250 highly differentiated SNPs were adjacent to at least one SV

Supplemental Table S6 SV frequency difference between representative cellulosic and sweet sorghum lines

Supplemental Table S7 Phenotypes used in analysis

Supplemental Table S8 Summary of insertions in putative transposon hotspots in the five genomes

Supplemental Table S9 Repeats annotation of the detected insertions in the putative transposon hotspots

Supplemental Table S10 Number of significant signals detected in GWAS based on SNP, SV and SNP+SV datasets

Supplemental Table S11 GWAS for 23 sorghum type related traits and 6 photoperiod related traits based on SNP, SV and SNP+SV datasets

Supplemental Table S12 Candidate genes underlying the correlated domestication of photoperiod sensitivity and sorghum type

Supplemental Table S13 DEGs as percentage of all genes in SV-associated gene and non-SV-associated gene categories in each annotation/effect fields

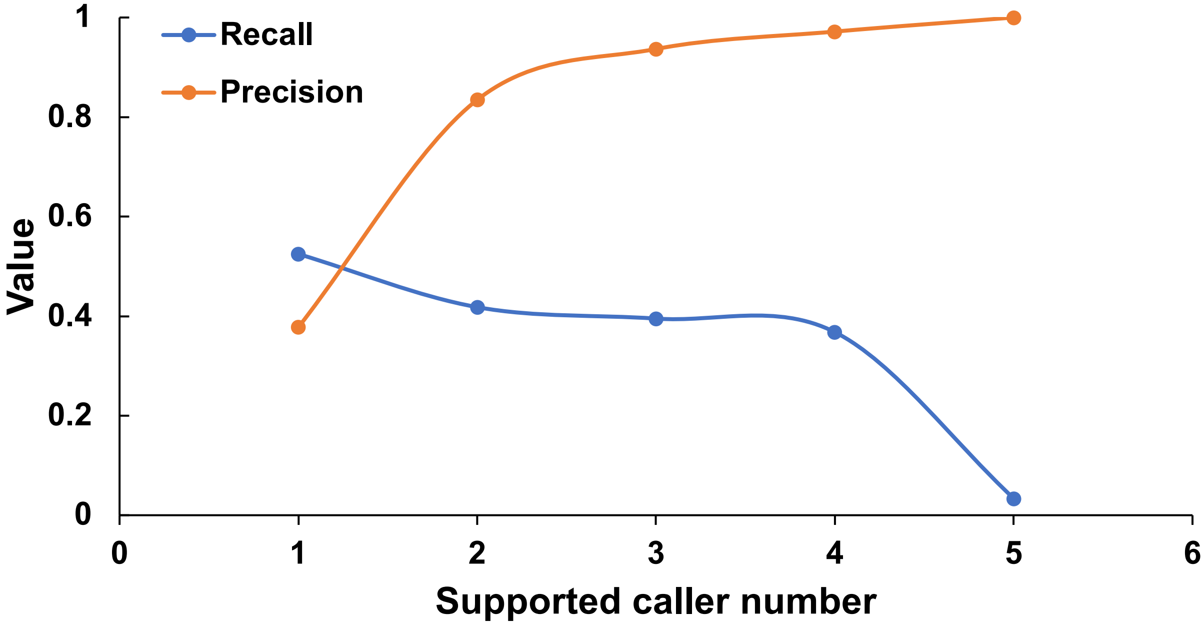
Supplemental Table S14 331 sorghum oil orthologs

Supplemental Table S15 Sorghum oil orthologs associated with CNV type SVs

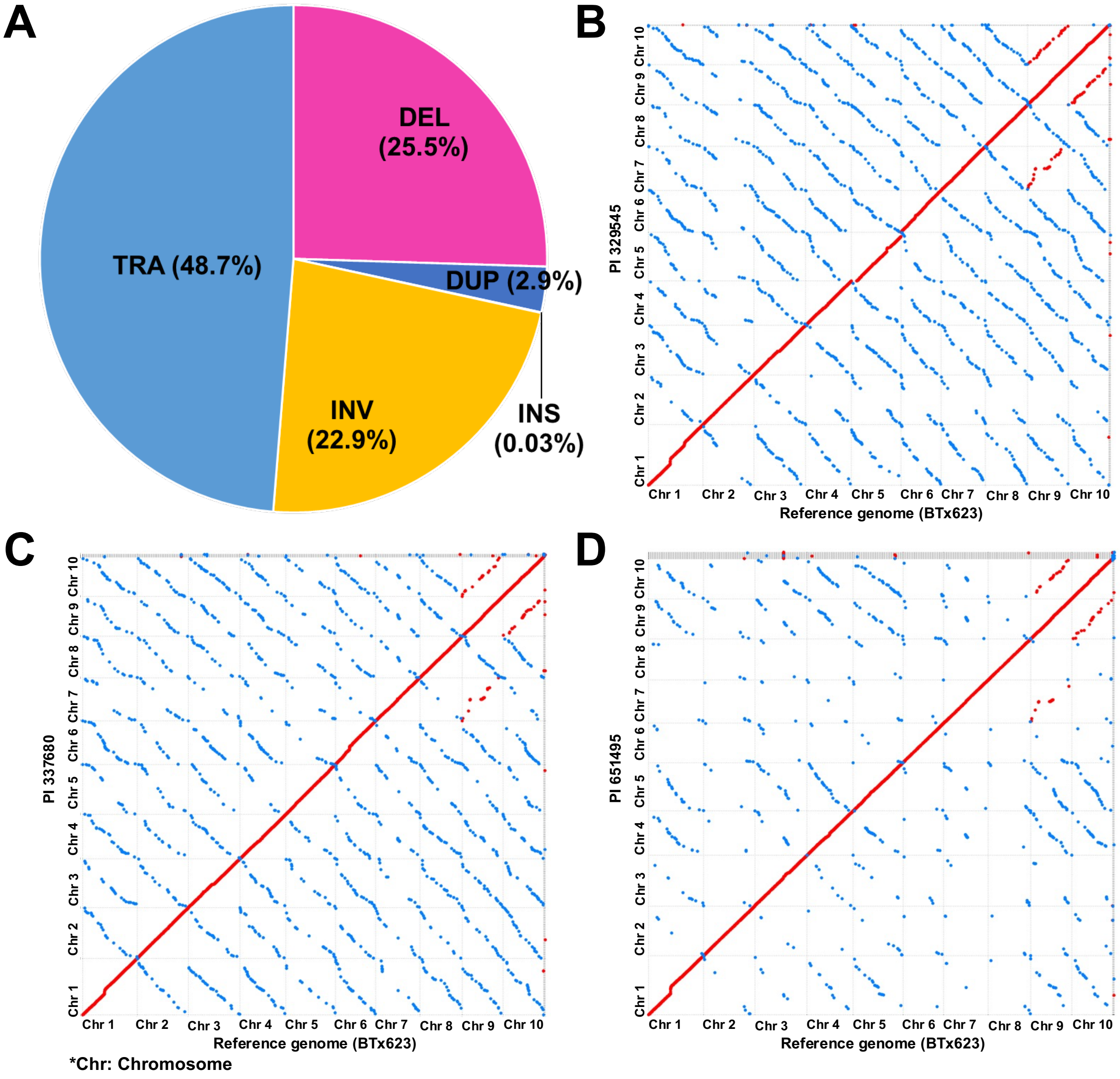
Supplemental Table S16 Sorghum oil orthologs associated with rearrangement type SVs

Supplemental Table S17 Precision, recall, F1, and Accuracy scores of the prediction model

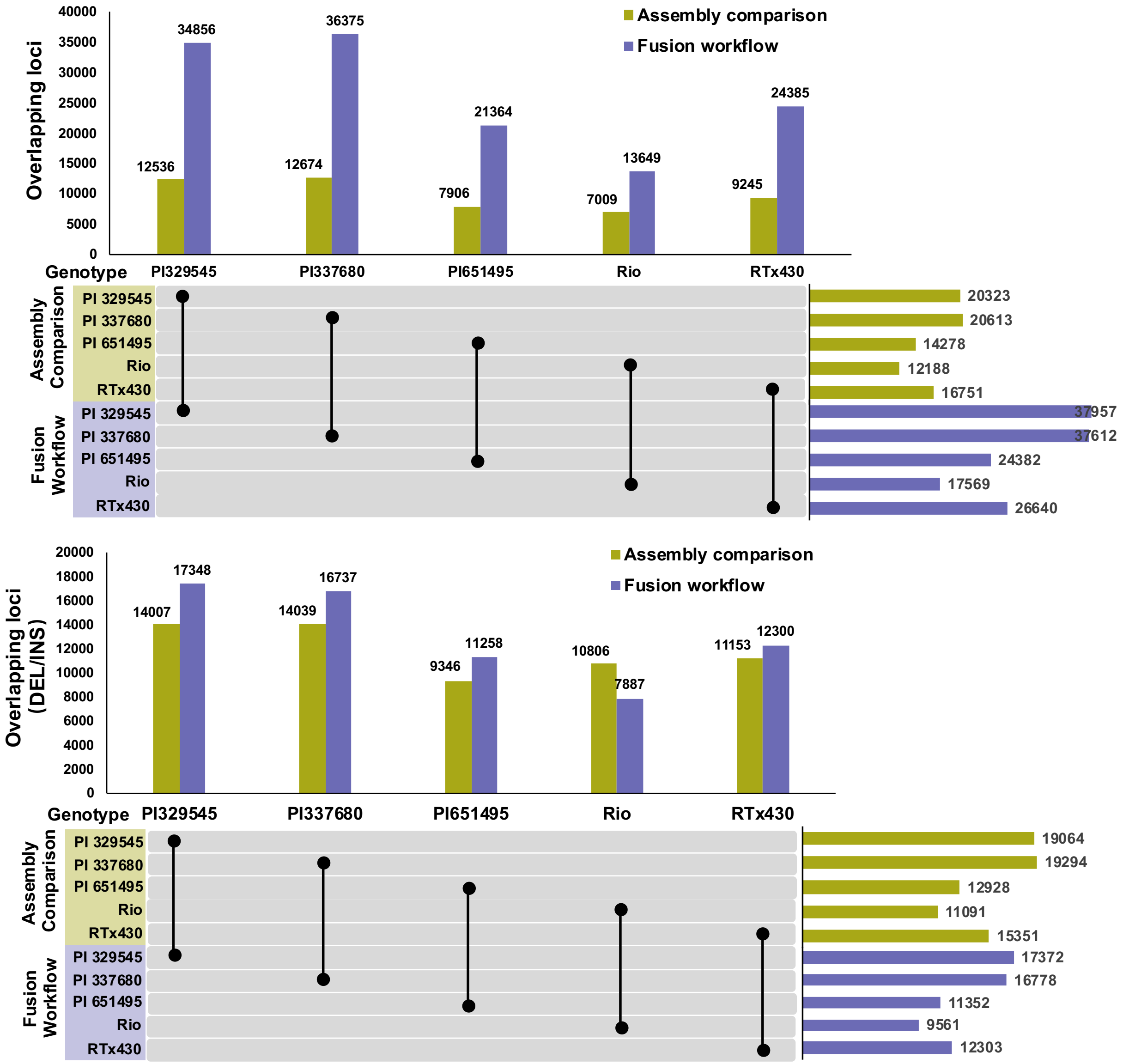
Supplemental Figures



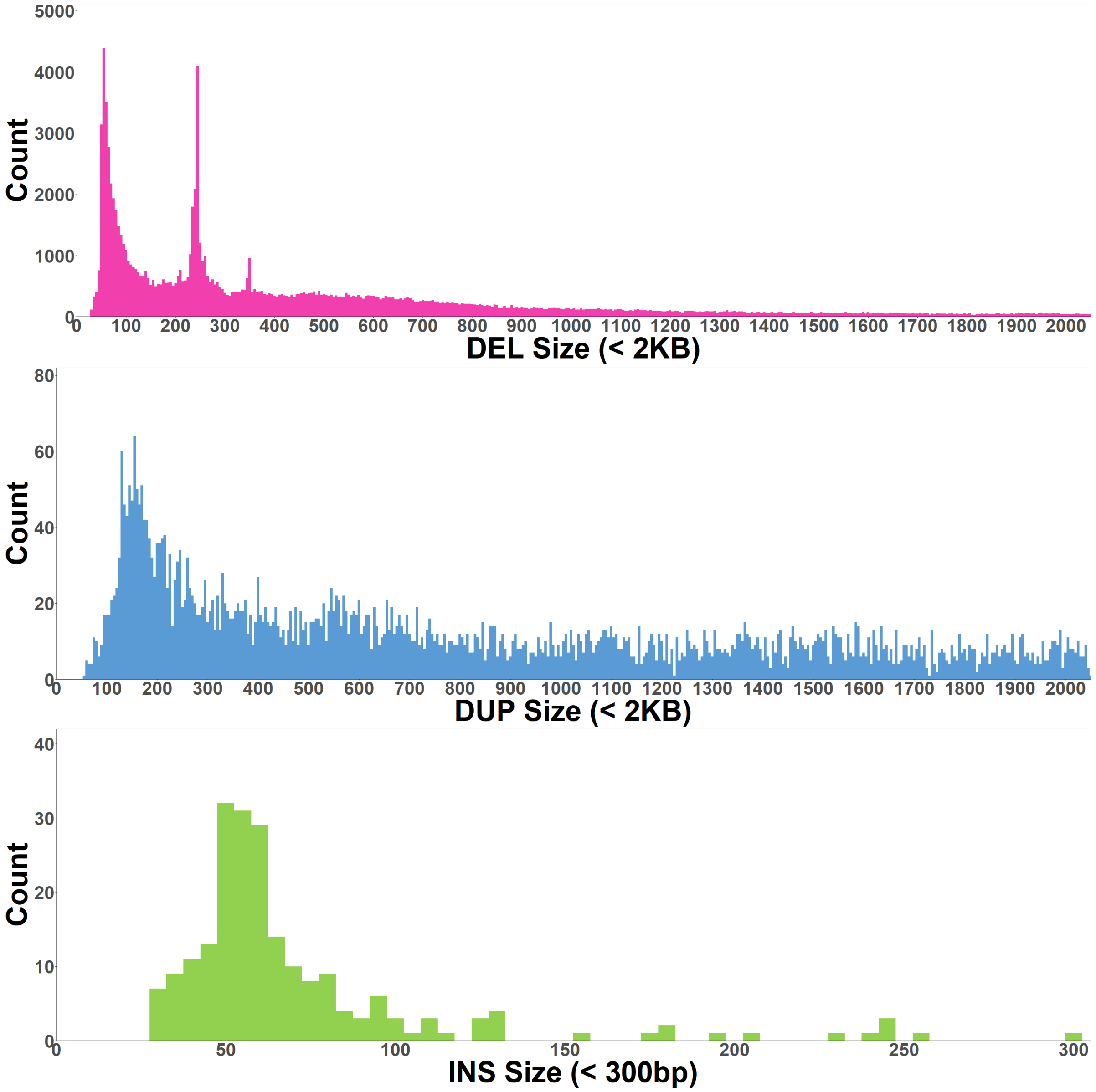
Supplemental Fig S1 Recall and precision values in different supporting callers. The precision rate increases while the recall rate decreases with the increment of the supported callers. The recall and precision trendlines cross between the "supported by at least 1 caller" and "supported by at least 2 callers" points.



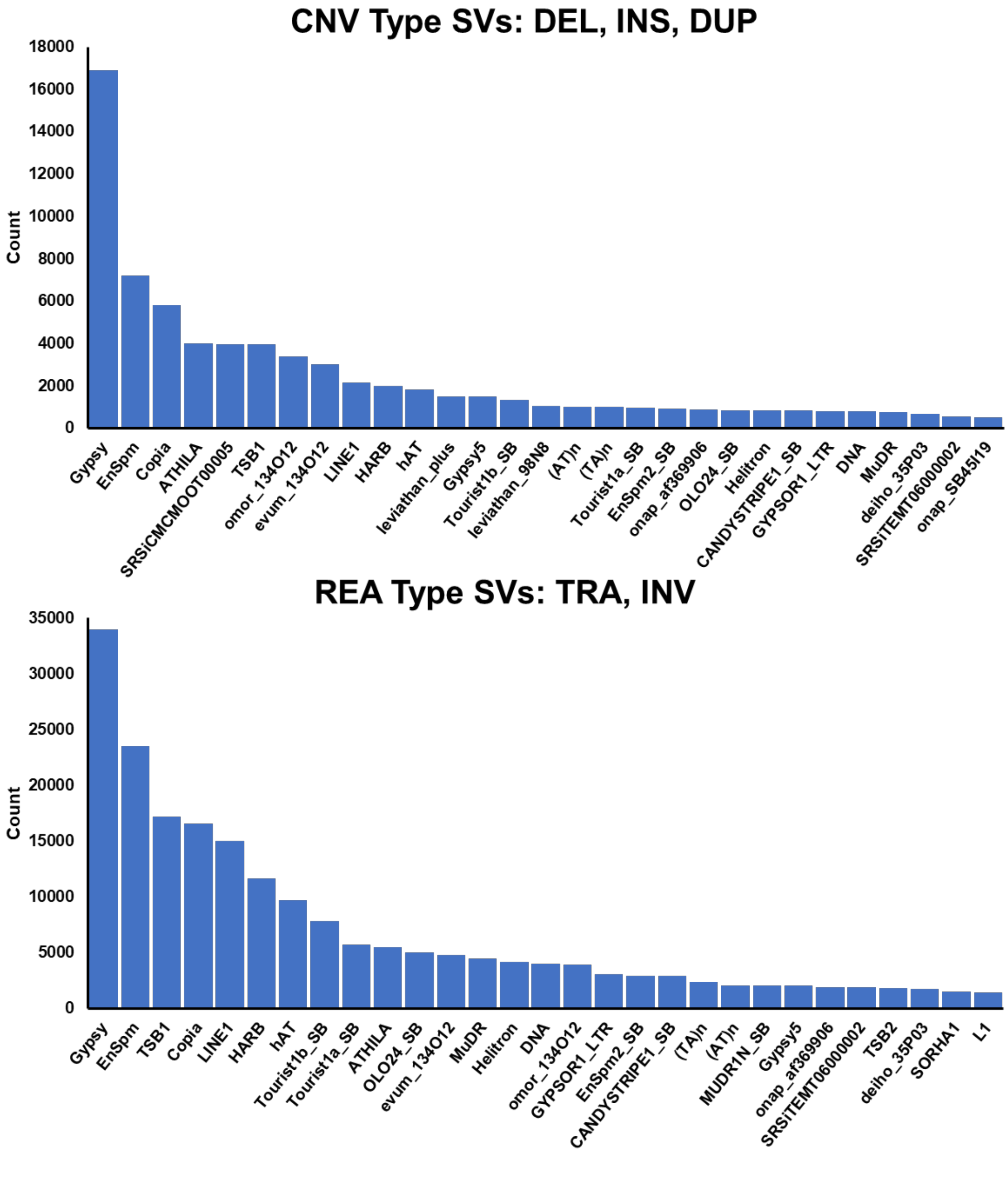
Supplemental Fig S2 Components of genome-wide structural variations (SVs) and evaluation. **A** Pie charts for the numbers of different SV types. There were 622,236 SVs identified on 10 chromosomes, including 158,614 deletions (DEL, 25.5%), 18,028 duplications (DUP, 2.9%), 216 insertions (INS, 0.03%), 142,219 inversions (INV, 22.9%) and 303,159 translocations (TRA, 48.7%). **B-D** mummer plots of the three assembled sorghum genomes, PI 329545 (**B**), PI 337680 (**C**), PI 651495 (**D**). The horizontal axis stands for the BTx623 reference, and the vertical axis stands for the three assembled genome.



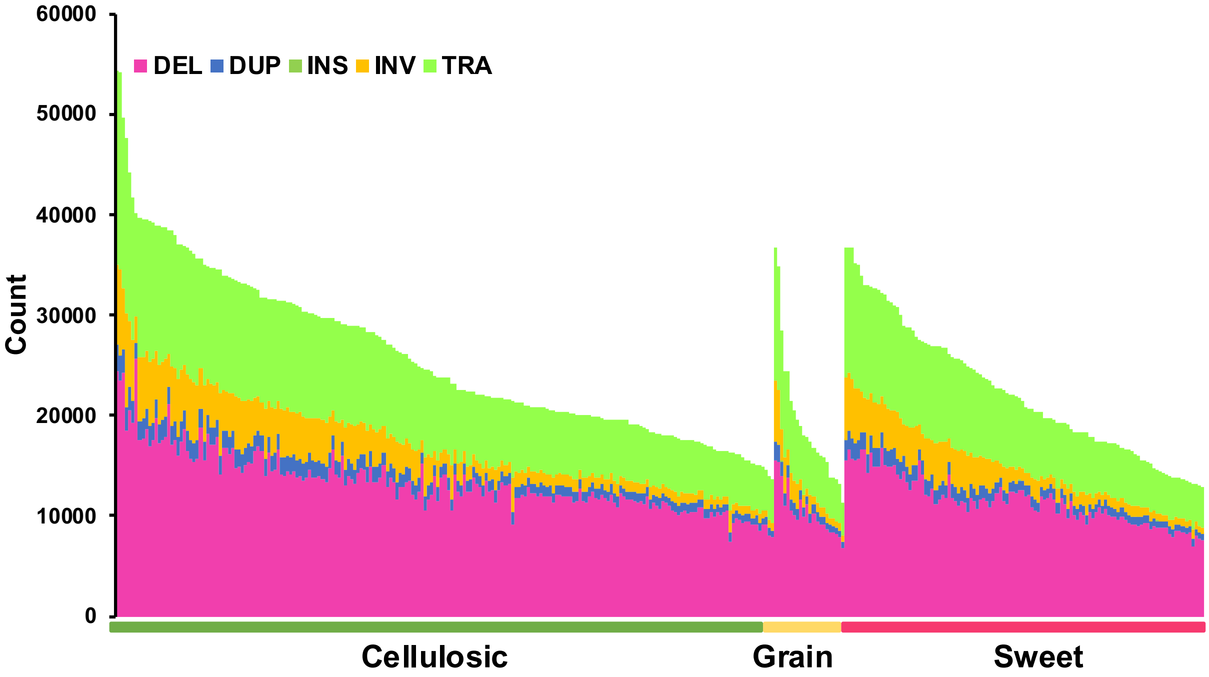
Supplemental Fig S3. Cross-validation of structural variants called at overlapping positions by assembly comparison and the fusion workflow. The upper plot shows the overlapping sets for all SVs; the lower plot shows data for only deletions and insertions. As the assembly comparison method, MUM&Co, tends to call large PAV polymorphisms, but fewer of them, there are fewer total polymorphisms but they encompass most variants called by the fusion method.



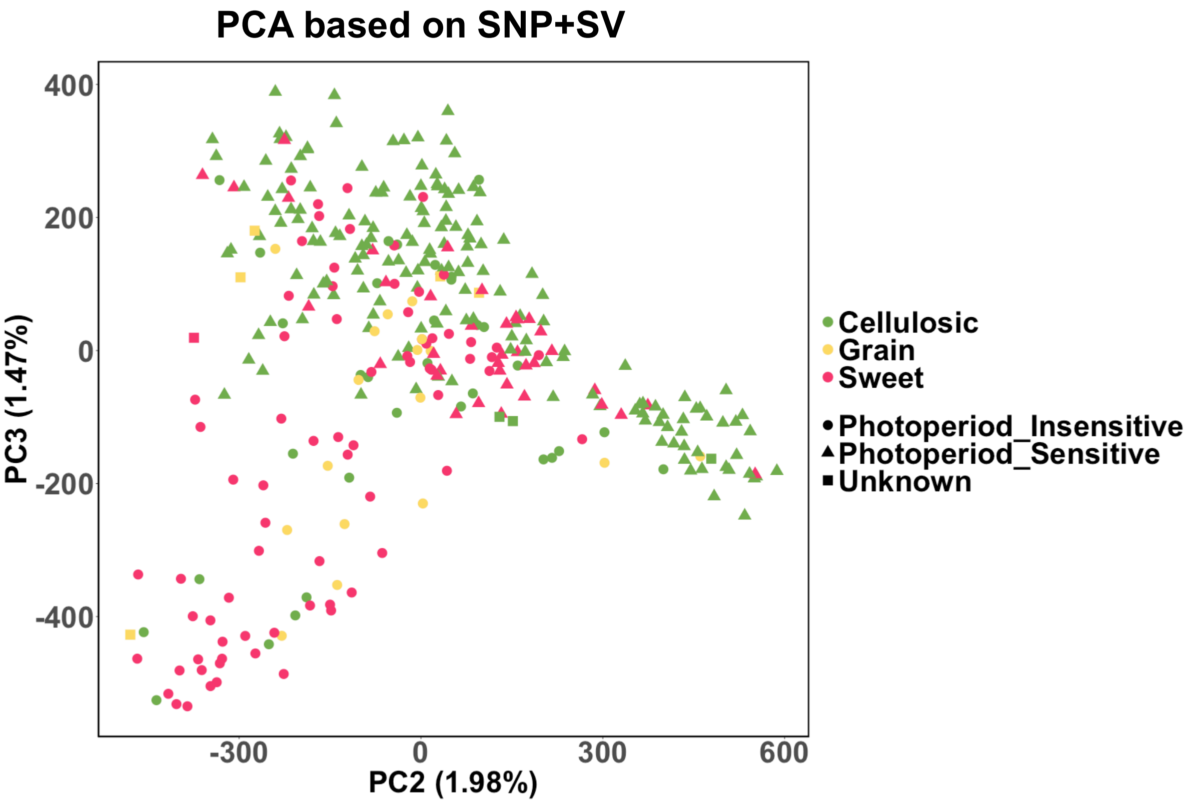
Supplemental Fig S4 Length distribution of copy number variant (CNV) type structural variations (SVs). Histogram of length distribution for CNV-type SVs. DEL (< 2KB), top; DUP (< 2KB), middle; INS (< 300bp), bottom. Most CNV-type SVs were relatively small: 30-250bp: 30.3%; 250-500bp: 13.1%; 500bp-1kb: 13.9%; 1kb-2kb: 9%; > 2kb: 33.6%.



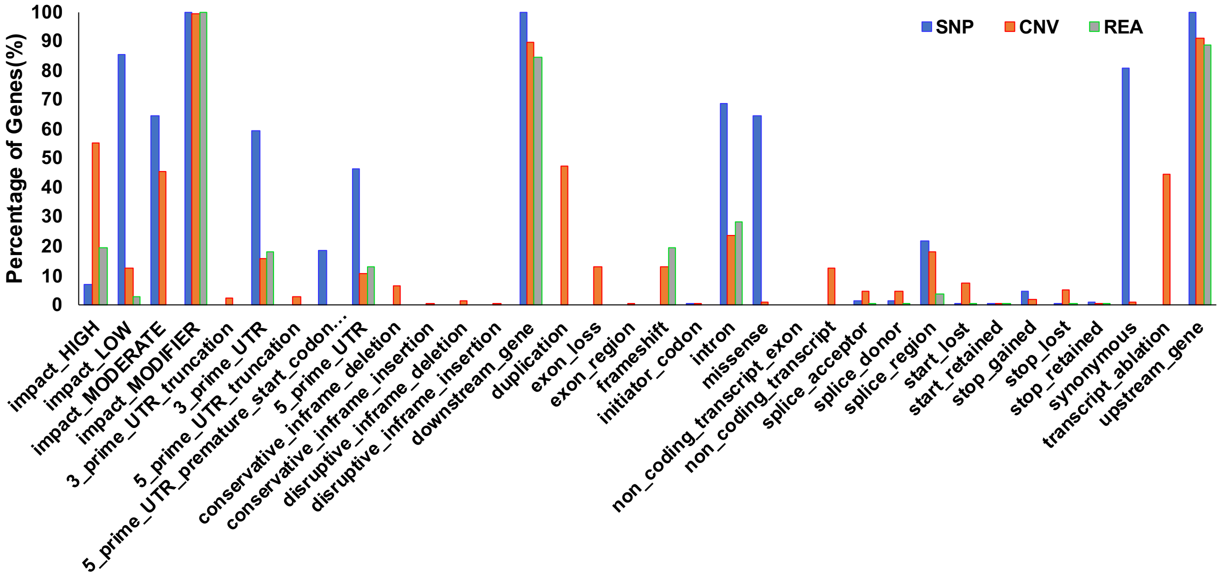
Supplemental Fig S5 Mobile elementary identification. Enrichment of different kinds of mobile elementaries hit by identified SVs. CNV-type (top) and rearrangement (REA) type (bottom) categories were employed based on the properties of the SVs spanning to conduct transposable elements impact analysis. The top two abundant transposable elements were *Gypsy* and *Enspm* in both CNV-type and REA-type SVs.



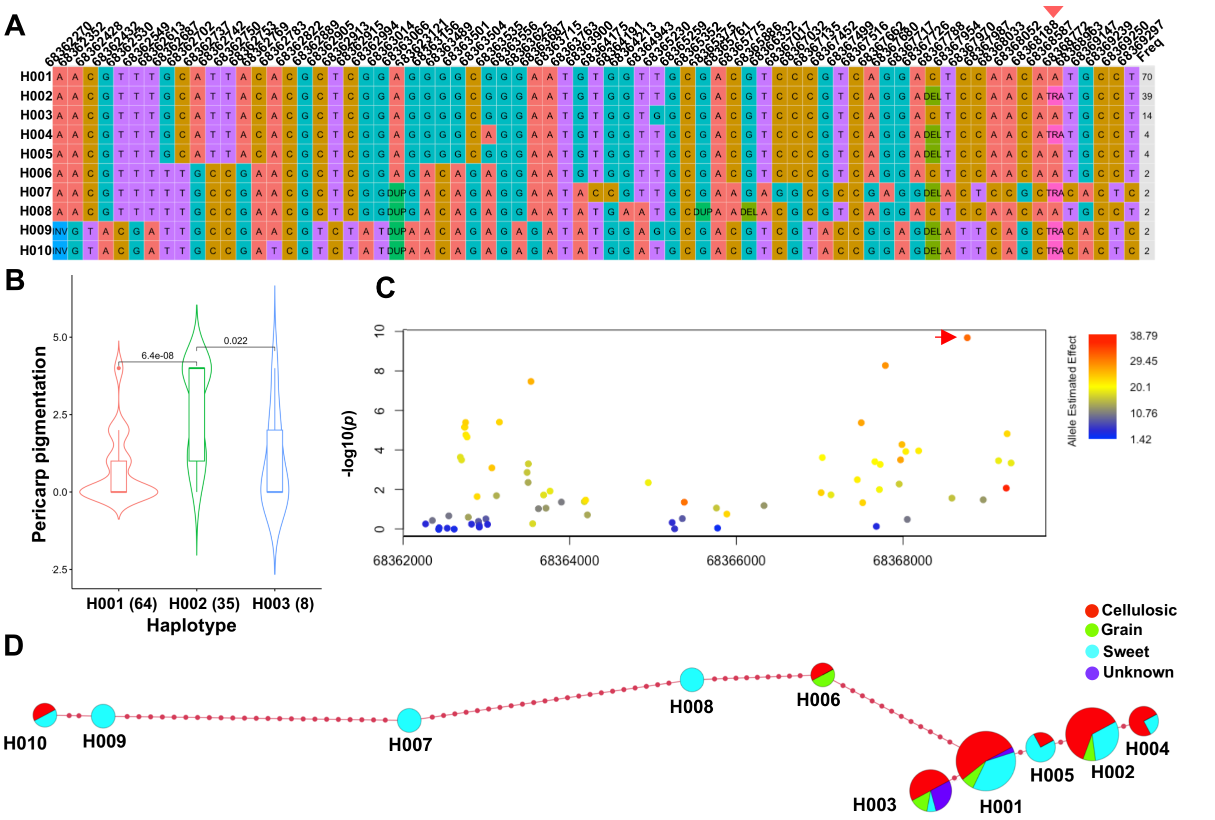
Supplemental Fig S6 Structural Variations **(**SVs) count distribution in different sorghum lines. The counts of SVs were sorted by total number of identified SVs within cellulosic (green), grain (yellow) and sweet (red) sorghum variety types in horizontal. SV compositions, including deletions (DEL, pink), duplications (DUP, blue), insertions (INS, green), inversions (INV, orange) and translocations (TRA, lime), were stacked by different colors columns individually. DEL and TRA are the major components of the total SVs for each accession. The sorghum containing the largest number of SVs was cellulosic type whereas the least one was grain type.



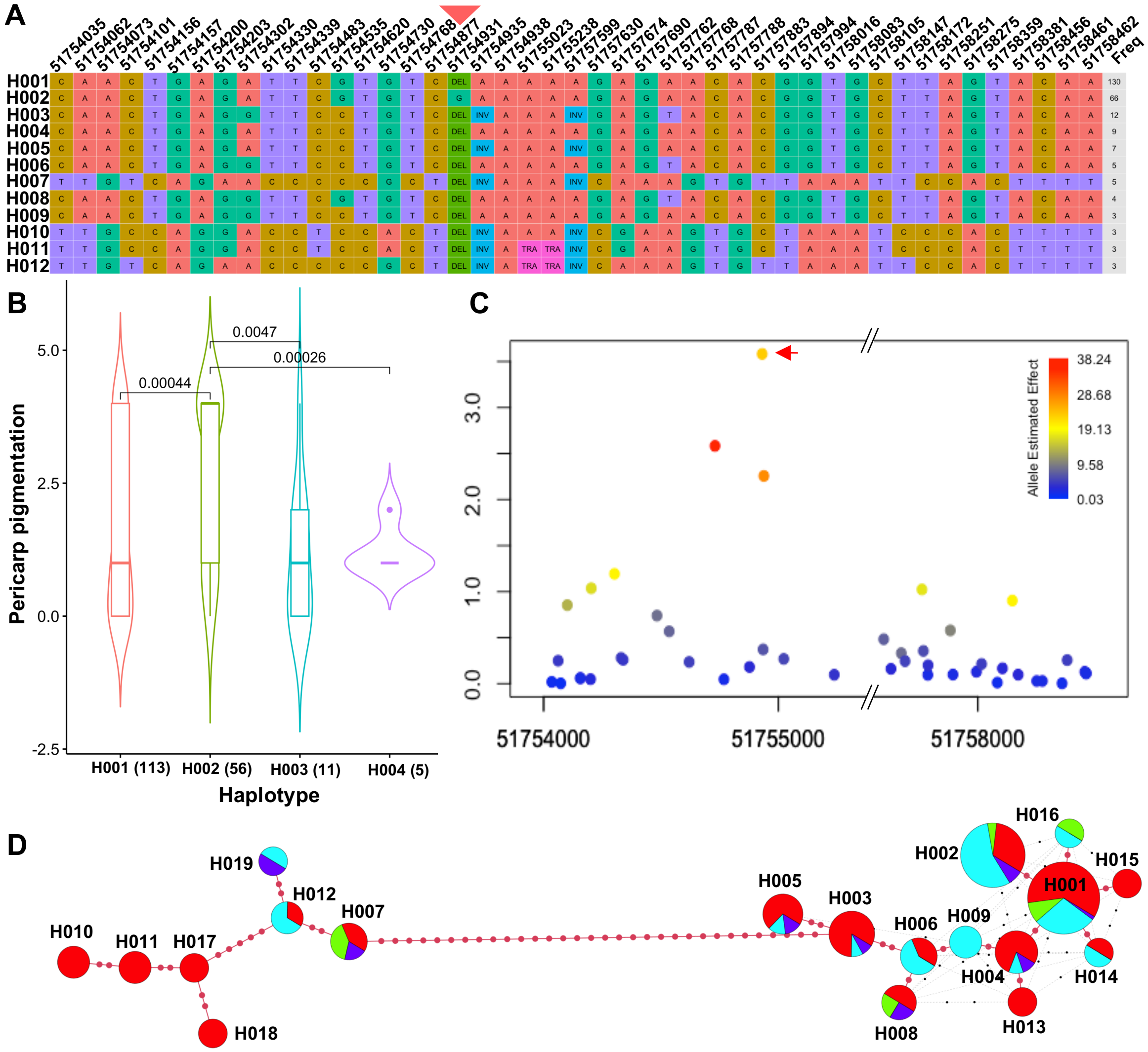
Supplemental Fig S7 Principal component analysis **(**PCA) based on single nucleotide polymorphism (SNP) + structural variations (SV) (right). Photoperiod sensitivity characters: Photoperiod_Insensitive (circle), Photoperiod_Sensitive (triangle) and unknow (square), and sorghum variety type information: cellulosic (green), grain (yellow) and sweet (red) were taken into consideration in PCA based on SNPs + SVs. Two obvious divided groups were derived from the population in all PCA results based on SNP, SV (**Figure 2B**), and SNP+SV datasets, even though sorghum lines in different variety types and photoperiod sensitivity have been undergoing frequent gene flows. The divided cellulosic sorghum was clustered with photoperiod sensitive character meanwhile the divided sweet sorghum was clustered with photoperiod insensitive feature.



Supplemental Fig S8 Annotation and functional effects of single nucleotide polymorphism (SNP) and structural variations (SVs) in sorghum genome. Percentages of genes impacted by different variants and annotation/effect types were calculated. Effects were categorized by impact levels firstly: “impact_High”, “impact_Moderate”, “impact_Low”, “impact_Modifier” which were pre-defined categories to summary the significance of variants. The following effect types in horizontal were sequence ontology effects, which provided the details for assessing sequence changes and impact. Copy number variant (CNV) type SVs and rearrangement (REA) type SVs were annotated separately. Blue bars, jacinth bars and grey bars with lime frame represent the percentages of genes affected by SNPs, CNV-type SVs and REA-type SVs in different effect types and sequence ontology effects respectively. SVs were generally responsible for high impacts, such as duplication, exon loss, codon frame shift and transcript ablation, whereas SNPs usually played a role in low impacts on genome including 3’ UTR, 5’ UTR, gene upstream and downstream regions, and intron sequences, which indicates a higher impact from SVs to sorghum genome than SNPs.

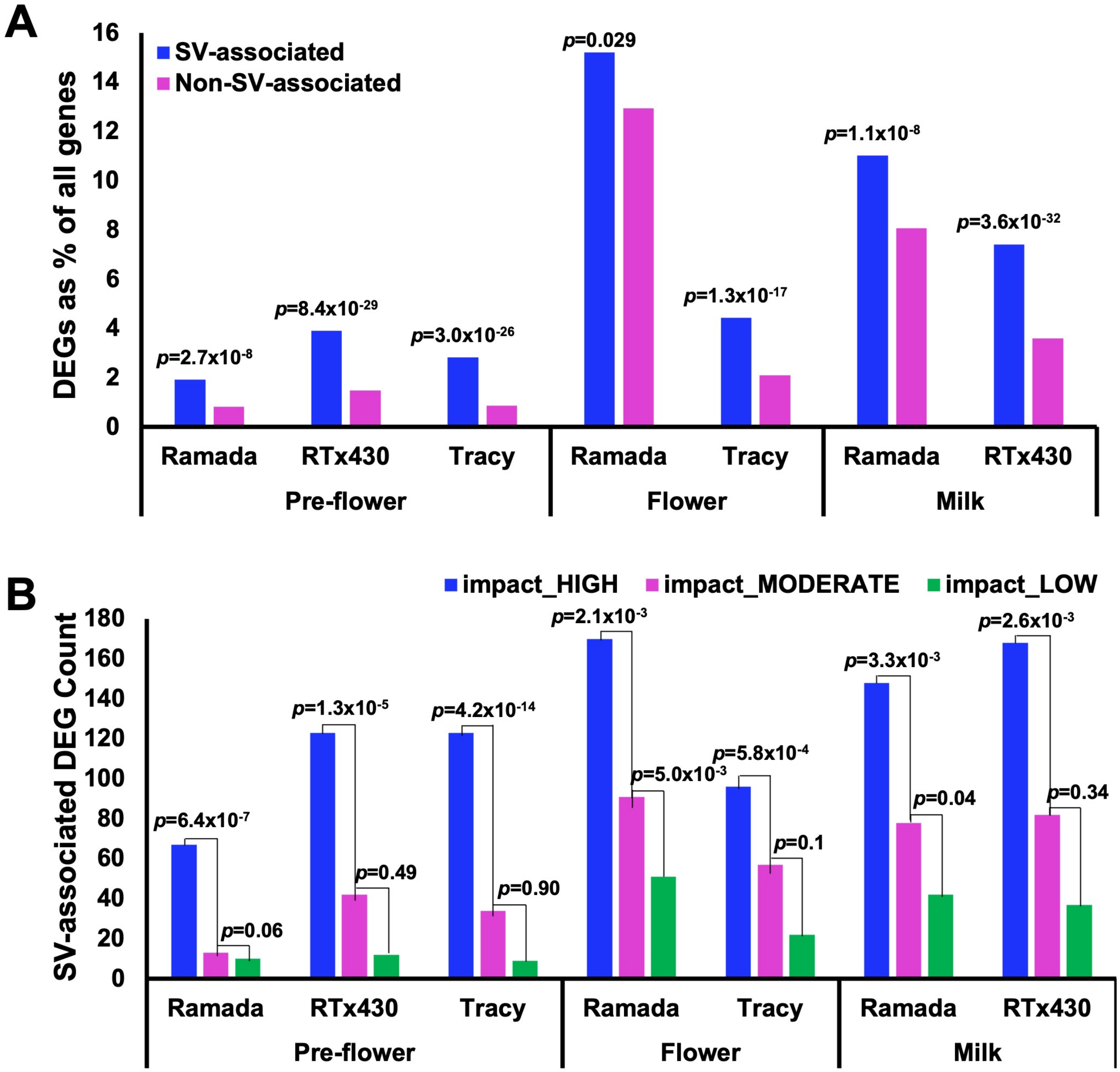


Supplemental Fig S9 Haplotype analyses of the TRA allele underlying *Y1* on Chromosome 1. **A** Haplotypes identified in BAP based on the variations surrounding the breakpoint of the TRA at 68,368,772 bp on Chromosome 1. Only the haplotypes with at least 2 sorghum lines are shown here. The red inverted triangle indicates the TRA position. **B** Phenotypic differences of three main haplotypes. Only the haplotypes containing at least five sorghum lines with available phenotype were compared. **C** Allele effect estimation analysis of the TRA allele. The red arrow indicates the TRA variant. **D** Haplotype network. Only the haplotypes with at least 2 sorghum lines are shown here.

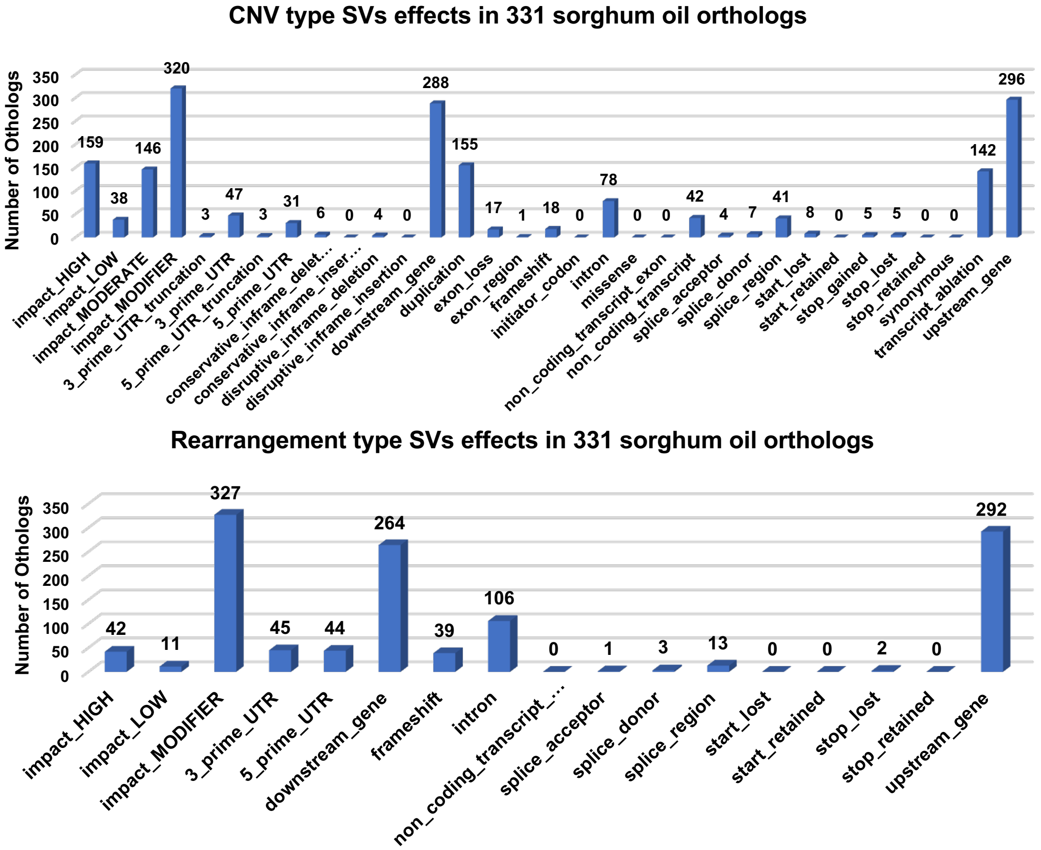


	51754035	51754062	51754073	51754101	51754156	51754157	51754200	51754203	51754302	51754330	51754339	51754483	51754535	51754620	51754730	51754768	51754877	51754931	51754935	51754938	51755023	51755238	51757599	51757630	51757674	51757690	51757762	51757768	51757787	51757788	51757883	51757894	51757994	51758016	51758083	51758105	51758147	51758172	51758251	51758275	51758359	51758381	51758456	51758461	51758462	Freq
H001	C	A	A	C	T	G	A	G	A	T	T	C	G	T	G	T	C	DEL	A	A	A	A	A	G	A	G	A	A	C	A	C	G	G	T	G	C	T	T	A	G	T	A	C	A	A	130
H002	C	A	A	C	T	G	A	G	A	T	T	C	G	T	G	T	C	G	A	A	A	A	A	G	A	G	A	A	C	A	C	G	G	T	G	C	T	T	A	G	T	A	C	A	A	66
H003	C	A	A	C	T	G	A	G	G	T	T	C	C	T	G	T	C	DEL	INV	A	A	A	INV	G	A	G	T	A	C	A	C	G	G	T	G	C	T	T	A	G	T	A	C	A	A	12
H004	C	A	A	C	T	G	A	G	A	T	T	C	C	T	G	T	C	DEL	A	A	A	A	A	G	A	G	A	A	C	A	C	G	G	T	G	C	T	T	A	G	T	A	C	A	A	9
H005	C	A	A	C	T	G	A	G	A	T	T	C	C	T	G	T	C	DEL	INV	A	A	A	INV	G	A	G	A	A	C	A	C	G	G	T	G	C	T	T	A	G	T	A	C	A	A	7
H006	C	A	A	C	T	G	A	G	G	T	T	C	C	T	G	T	C	DEL	A	A	A	A	A	G	A	G	T	A	C	A	C	G	G	T	G	C	T	T	A	G	T	A	C	A	A	5
H007	T	T	G	T	C	A	G	A	A	C	C	C	C	C	G	C	T	DEL	INV	A	A	A	INV	C	A	A	A	G	T	G	T	T	A	A	A	T	T	C	C	A	C	T	T	T	T	5
H008	C	A	A	C	T	G	A	G	G	T	T	C	G	T	G	T	C	DEL	A	A	A	A	A	G	A	G	T	A	C	A	C	G	G	T	G	C	T	T	A	G	T	A	C	A	A	4
H009	C	A	A	C	T	G	A	G	G	T	T	C	C	T	G	T	C	DEL	A	A	A	A	A	G	A	G	A	A	C	A	C	G	G	T	G	C	T	T	A	G	T	A	C	A	A	3
H010	T	T	G	C	C	A	G	G	A	C	C	T	C	C	A	C	T	DEL	INV	A	A	A	INV	C	G	A	A	G	T	G	C	T	A	A	A	T	C	C	C	A	C	T	T	T	T	3
H011	T	T	G	C	C	A	G	G	A	C	C	T	C	C	A	C	T	DEL	INV	A	TRA	TRA	INV	C	G	A	A	G	T	G	C	T	A	A	A	T	C	C	C	A	C	T	T	T	T	3
H012	T	T	G	T	C	A	G	A	A	C	C	C	C	C	G	C	T	DEL	INV	A	TRA	TRA	INV	C	A	A	A	G	T	G	T	T	A	A	A	T	T	C	C	A	C	T	T	T	T	3

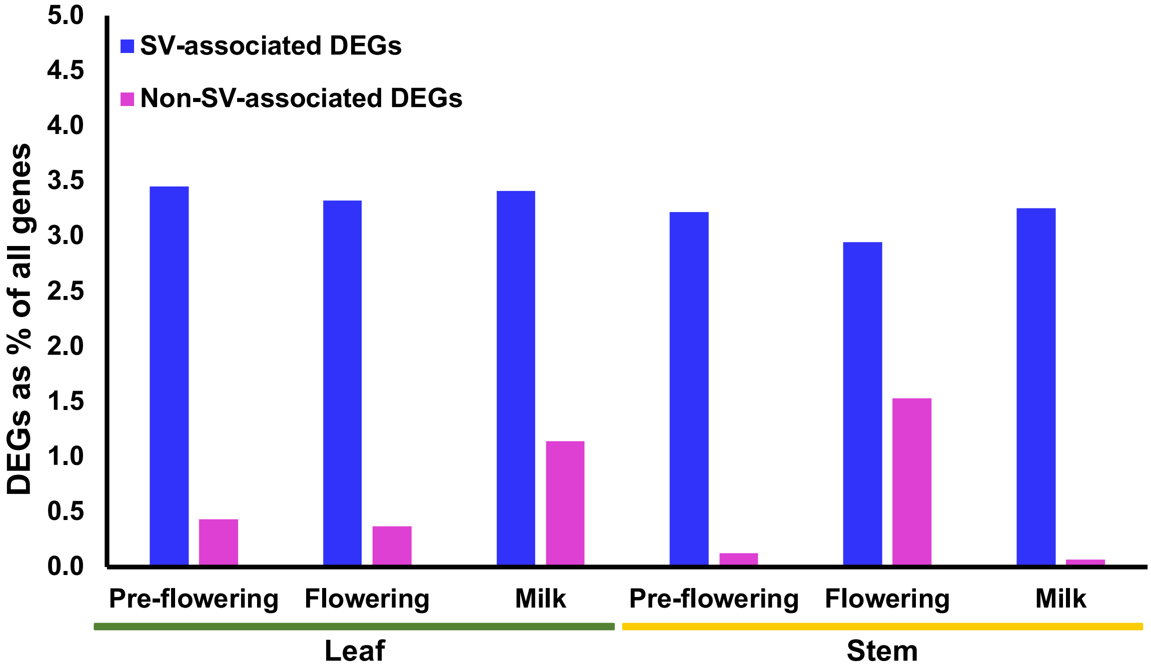
Supplemental Fig S10 Haplotype analyses of the 2.6 kb DEL on Chromosome 8. **A** Haplotypes identified in BAP based on the variations surrounding the 2.6 kb DEL. Only the haplotypes with at least 3 sorghum lines are shown here. The red inverted triangle indicates the DEL position. **B** Phenotypic differences of three main haplotypes. Only the haplotypes containing at least five sorghum lines with available phenotype were compared. **C** Allele effect estimation analysis of the DEL allele. The red arrow indicates the DEL variant. **D** Haplotype network. Only the haplotypes with at least 2 sorghum lines are shown here.



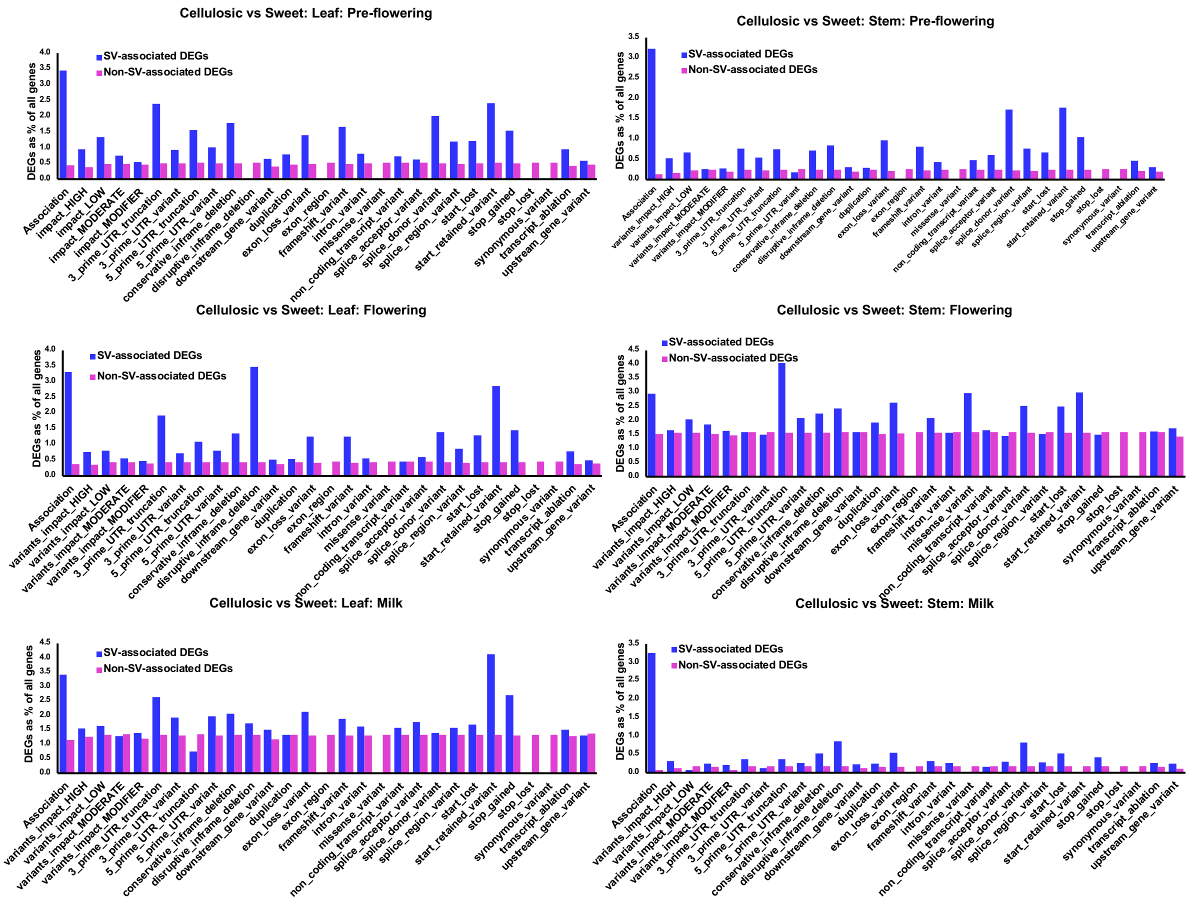
Supplemental Fig S11 Structural variants (SVs) have a widespread impact on gene expression. **A** SVs have an impact on gene expression in sorghum stem across all developmental stages. Differentially expressed gene (DEG) analysis was performed by comparison of expression profiles in RTx430, Tracy and Ramada with the expression profile in Tx623 (as control) in stem and three development stages. Blue and pink bars represent the SV-associated and non-SV-associated DEGs as percentages of all genes, respectively. The *p* values on the top of SV-associated DEG bars indicate the hypergeometric testing results for enrichment of DEGs in SV-associated genes. The *p* values were adjusted using Bonferroni correction. DEGs were significantly enriched in SV-associated genes. **B** SV-associated DEG count changed according to different impact predictions. The vertical axis showed the SV-associated DEG count. Blue, pink and green bars represent the DEG counts associated by high impact SVs (impact_HIGH), moderate impact SVs (impact_MODERATE) and low impact SVs (impact_LOW) respectively in leaf tissue of different sorghum lines in three developmental stages (pre-flower, flower and milk). The *p* values show the significance levels between groups (see Methods). Differential DEG counts between "impact_HIGH" and "impact_MODERATE" were all statistically significant. Significant level of DEG counts between "impact_MODERATE" and "impact_LOW" varied depending on lines and stages. In general, higher impact SVs associated more DEGs.



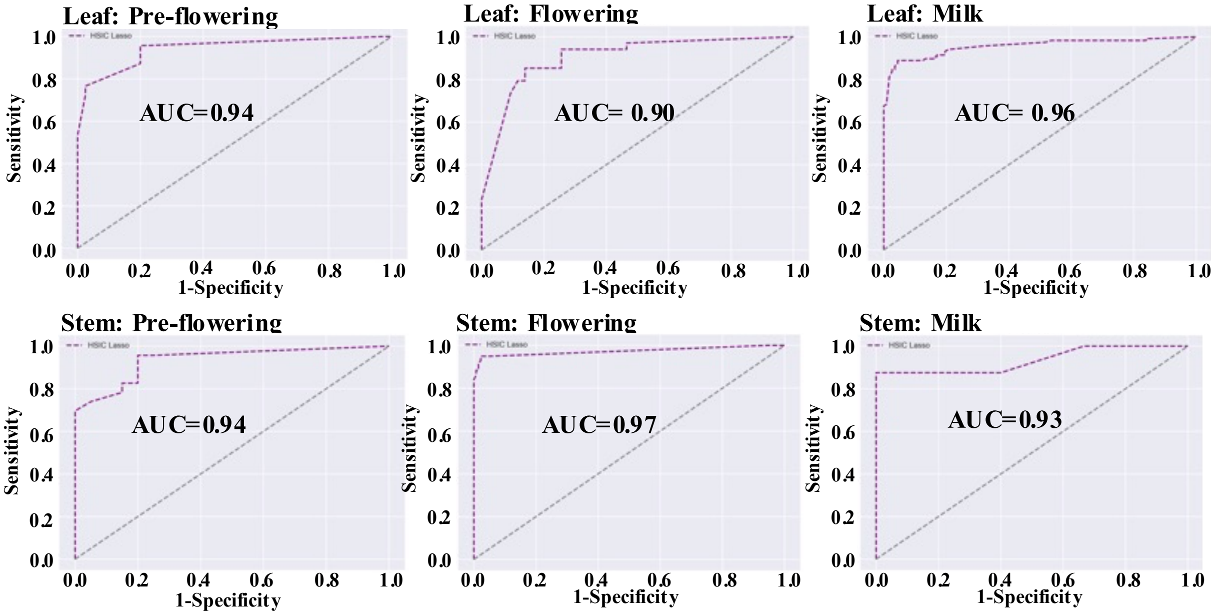
Supplemental Fig S12 Different functional effects of SVs on 331 sorghum oil orthologs. Different functional effects of CNV type SVs on 331 sorghum oil orthologs (upper). Different functional effects of rearrangement type SVs on 331 sorghum oil orthologs (lower).



Supplemental Fig S13 Structural variants (SVs) have an impact on gene expression in other cellulosic-sweet sorghum comparison beyond BAP. Differentially expressed gene (DEG) analysis was performed by comparison of expression profiles in cellulosic group and sweet sorghum group beyond BAP in leaf and stem tissues, and three development stages. Blue and pink bars represent the SV-associated and non-SV-associated DEGs as percentages of all genes, respectively. The SV-associated DEGs as percentage of all genes are still much higher than non-SV-associated DEGs as percentage of all genes in all stages and tissues.



Supplemental Fig S14 Different functional effects of SVs on differentially expressed genes (DEGs) in other cellulosic-sweet sorghum comparison beyond BAP. CNV type variations specific to the representative cellulosic or sweet sorghum lines were deployed to predict the gene expression in additional cellulosic and sweet accessions in 3 development stages: pre-flowering, flowering, and milk, and 2 tissues: leaf and stem. Blue and pink bars represent the SV-associated and non-SV-associated DEGs as percentages of all genes, respectively.



Supplemental Fig S15 AUCROC scores of prediction model based on Block Hilbert Schmidt Independence Criterion Lasso (Block HSIC Lasso) method across different tissues and stages. The AUCROC scores show a possibility of the prediction from SVs to differentiated gene expression.