frontiers in Plant Science

ORIGINAL RESEARCH
published: 01 June 2021
doi: 10.3389/fpls.2021.666342

Sorghum Pan-Genome Explores the Functional Utility for Genomic-Assisted Breeding to Accelerate the Genetic Gain

Pradeep Ruperao[1], Nepolean Thirunavukkarasu[2], Prasad Gandham[1], Sivasubramani Selvanayagam[1], Mahalingam Govindaraj[1], Baloua Nebie[3], Eric Manyasa[4], Rajeev Gupta[1], Roma Rani Das[1], Damaris A. Odeny[4], Harish Gandhi[1], David Edwards[5], Santosh P. Deshpande[1] and Abhishek Rathore[1]**

[1] International Crops Research Institute for the Semi-Arid Tropics, Patancheru, India, [2] Genomics and Molecular Breeding Lab, ICAR-Indian Institute of Millets Research, Hyderabad, India, [3] Sorghum Breeding Program, International Crops Research Institute for the Semi-Arid Tropics, Bamako, Mali, [4] Sorghum Breeding Program, International Crops Research Institute for the Semi-Arid Tropics, Nairobi, Kenya, [5] School of Biological Sciences and Institute of Agriculture, The University of Western Australia, Perth, WA, Australia

OPEN ACCESS

Edited by:
Sean Mayes,
University of Nottingham,
United Kingdom

Reviewed by:
Bernardo Ordas,
Consejo Superior de Investigaciones Científicas (CSIC), Spain
Fernando Martinez,
Sevilla University, Spain

***Correspondence:**
Abhishek Rathore
a.rathore@cgiar.org
Santosh P. Deshpande
s.deshpande@cgiar.org

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: 18 February 2021
Accepted: 28 April 2021
Published: 01 June 2021

Citation:
Ruperao P, Thirunavukkarasu N, Gandham P, Selvanayagam S, Govindaraj M, Nebie B, Manyasa E, Gupta R, Das RR, Odeny DA, Gandhi H, Edwards D, Deshpande SP and Rathore A (2021) Sorghum Pan-Genome Explores the Functional Utility for Genomic-Assisted Breeding to Accelerate the Genetic Gain. Front. Plant Sci. 12:666342. doi: 10.3389/fpls.2021.666342

Sorghum (*Sorghum bicolor* L.) is a staple food crops in the arid and rainfed production ecologies. Sorghum plays a critical role in resilient farming and is projected as a smart crop to overcome the food and nutritional insecurity in the developing world. The development and characterisation of the sorghum pan-genome will provide insight into genome diversity and functionality, supporting sorghum improvement. We built a sorghum pan-genome using reference genomes as well as 354 genetically diverse sorghum accessions belonging to different races. We explored the structural and functional characteristics of the pan-genome and explain its utility in supporting genetic gain. The newly-developed pan-genome has a total of 35,719 genes, a core genome of 16,821 genes and an average of 32,795 genes in each cultivar. The variable genes are enriched with environment responsive genes and classify the sorghum accessions according to their race. We show that 53% of genes display presence-absence variation, and some of these variable genes are predicted to be functionally associated with drought adaptation traits. Using more than two million SNPs from the pan-genome, association analysis identified 398 SNPs significantly associated with important agronomic traits, of which, 92 were in genes. Drought gene expression analysis identified 1,788 genes that are functionally linked to different conditions, of which 79 were absent from the reference genome assembly. This study provides comprehensive genomic diversity resources in sorghum which can be used in genome assisted crop improvement.

Keywords: sorghum, pan-genome, diversity, SNP, gPAV, GWAS, drought genes

INTRODUCTION

Sorghum (*Sorghum bicolor*) is a multi-utility cereal of global importance, and a major food crop in sub-Saharan Africa and South Asia (Ritter et al., 2007; Motlhaodi et al., 2014). It is typically a diploid species ($2n = 20$) with an estimated genome size of the ~800 Mb sequence (Price et al., 2005). It provides important primary and secondary products, such as food, fodder, starch, fibre,

biofuels, alcohol, dextrose syrup as well as other products. It is domesticated and further bred for diverse use as food, fodder, and bioenergy in different agro-climatic conditions (Li et al., 2010) and shows a wide diversity at the genome level (Kong et al., 2000; Hart et al., 2001).

A draught sorghum genome assembly of 730 Mb was initially prepared for *Sorghum bicolor* Moench (Paterson et al., 2009), followed by an improved assembly of 732.2 Mb, covering ∼91.5% of the genome (McCormick et al., 2018). Recently, a sorghum reference genome assembly for the "Rio" line was generated comprising 729 Mb (Cooper et al., 2019). Each of these genome assemblies is limited to its respective accession and does not reflect the diversity of genes in this species.

The presence or absence of genes or genomic regions among genotypes is an important form of genomic variation in plants, and genes can be categorised into core and variable within the species (Saxena et al., 2014; Golicz et al., 2016b). The collection of these core and variable genes is known as pan-genome. Studying the pan-genome from a large number of genotypes enhances the understanding of species diversity, domestication and breeding history, and provides complete characterisation of species genes content diversity as demonstrated in rice (Wang et al., 2018) and tomato (Gao et al., 2019).

Several approaches are available to construct a pan-genome (Golicz et al., 2016b). The classical approach of whole-genome assembly of all genotypes was initially implemented in bacteria, and later developments led to the complementary method to "iteratively map and assemble," the unmapped sequence reads, demonstrated in *B. oleracea* (Golicz et al., 2016a), *B. napus* (Hurgobin et al., 2018), bread wheat (Montenegro et al., 2017), and pigeon pea (Zhao et al., 2020). The whole genome assembly and comparison method has the advantage in that it can place almost all individual specific genes in a genomic context, but suffers from the inability to distinguish assembly or annotation errors from true biological variation (Bayer et al., 2017). It is also unsuitable for large population studies due to the expense of sequencing, assembling, and comparing large numbers of genomes. In contrast, the iterative assembly approach can cost effectively assess large numbers of genotypes for gene presence/absence variation and hence identify genes that may be relatively rare in a population and not samples in whole genome assembly approaches, though without additional long read data, it is unable to place many of the newly identified genes. Hence the iterative assembly method is most suited for large population diversity studies.

Hence, we assembled a pan-genome using reference and re-sequenced genomes for genetically diverse race-specific sorghum accessions. The sorghum pan-genome was initiated with the reference genome obtained from JGI on Phytozome (McCormick et al., 2018), followed by adding to this reference with novel genome sequences from 176 sorghum accessions.

We provided structural and potential functional aspects of this pan-genome in the form of genes, single nucleotide polymorphism (SNP) and gene presence and absence variations (PAV). The utility of the pan-genome was demonstrated by identifying candidate functional genes using publicly available SNP chip data, genome-wide association studies and gene-expression assays. These sorghum pan-genome resources will be useful for achieving the sustainable development goals in developing countries by accelerating the genetic gain in arid and semi-arid ecologies.

MATERIALS AND METHODS

Pan-Genome Assembly and Annotation

The pan-genome was assembled using iterative mapping and assembly approaches. The assembly was initiated with a sorghum reference assembly v3.0.1 to map sorghum accessions whole-genome sequence data iteratively. Reads from 176 sorghum accessions with a minimum of 10X coverage sequence data were mapped to the sorghum reference v3.0 (McCormick et al., 2018) using Bowtie2 (Langmead and Salzberg, 2012) v2.3.4, and unmapped reads were assembled with IDBA_UD assembler (Peng et al., 2012) and the assembled contig sequence more than 500 bp length was only considered and appended to reference genome sequence. The resulting final assembly sequence was compared with NCBI non-redundant nucleotide databases using BLASTn and the sequences with homology to sorghum mitochondria (NC_008360), chloroplast (MK348612) also the sequences having homology outside the green plant group Viridiplantae taxonomy group (Taxonomy ID: 33090) were removed. The remaining sequences were self-compared with nucmer search (http://mummer.sourceforge.net/) and sequences with >90 percent coverage with greater 90 percent identity were removed to maintain the non-redundancy of the novel sequences. REPEATMASKER-v4.0.7 (Smit et al., 2000) masked repetitive elements using sorghum as the species. The sorghum expressed sequence tags (ESTs) from GenBank were aligned with tBLASTx and genes were predicted using AUGUSTUS v3.3.2, supporting the EST alignments. The gene models having fewer than 300 bp in length were filtered out and the remaining genes supporting either EST alignments or hisat2 (Kim et al., 2019) alignments (RNASeq read from 25 accessions, **Supplementary Table 1**) further used for functional annotation against uniref90 (database downloaded in May 2020).

Gene Presence-Absence Variations (gPAVs)

Whole-genome sequence reads of all 354 sorghum accessions were mapped with Bowtie2 v2.3.4 (Langmead and Salzberg, 2012) to pan-genome assembly with a wide insert size range between 0 and 1,000 bp. The gene PAVs were defined based on sequence reads coverage mapped to respective genes as described by Golicz et al. (2016a). Genes models on contigs longer than 1 Kbp were used in this analysis. PAV converted into the binary matrix and with 1,000 bootstrap resampling were used to estimate the genetic relationship among the accessions with R "ape" package (Paradis et al., 2004) to construct an NJ tree and visualised in iTOL tree viewer (Letunic and Bork, 2019).

The core genes were defined as the genes present in all the accessions, whereas the variable genes are the genes missing in one or more accessions. The *in-house* developed script was used to define the core and variable genes from the PAV matrix. Core and variable genes were compared for gene length, exon number,

synonymous SNPs, non-synonymous SNPs, and Ka/Ks. The mean count for each sample size of core and pan-genes present in all possible combinations of 354 accessions was plotted. The protein sequences of *Zea mays, Setaria italica, Brachypodium distachyon,* and *Oryza sativa* were downloaded from the public database UniProt for cluster analysis. All protein sequences were compared using all-by-all BLASTp followed by MCL for gene clustering into gene families with default parameters. The gene enrichment analysis was performed with Fisher exact test from R "topGO" package (Alexa et al., 2006) using "Elim" method.

SNP Discovery and Annotation

The sorghum whole genome sequence reads of 354 accessions were quality trimmed using Trimmomatic (Bolger et al., 2014) and mapped to pan-genome using Bowtie2 v2.3.4 (Langmead and Salzberg, 2012) allowing to map paired reads. The aligned reads in SAM format converted to BAM format using samtools (Li et al., 2009) followed by filtering out the read duplication with Picard tools (http://broadinstitute.github.io/picard). Variants against the reference (pan-genome) were called with GATK v.4.1 (McKenna et al., 2010) and directed to quality filtered with vcftools v.0.1.13 (Danecek et al., 2011). The variant sites having missing genotypes of more than 0.15 and minor allele count <2 were excluded and the remaining sites were used for downstream analysis such as SNP functionally annotated with SnpEff v.4.3 (Cingolani et al., 2012).

Sorghum Diversity and Population Structure

A subset of 216 diverse sorghum accessions from 354 set with known sorghum race information (Valluru et al., 2019), was used for genetic diversity and population structure assessment. A total of 1.12 million filtered SNPs from sorghum race accessions were retained for downstream analysis. The STRUCTURE v2.3 (Hubisz et al., 2009), was used to estimate the population structure using the admixture model. The tested K was set from 2 to 5 and optimal K for population structure was defined with the structure program. With the same SNP set, PCo analysis was done with R labdsv package (https://CRAN.R-project.org/package=labdsv) and phylogeny analysis performed using 1000 replicates with R "ape" package (Paradis et al., 2004) and visualised in iTOL tree viewer (Letunic and Bork, 2019).

Genome-Wide Association Analysis (GWAS)

Two different mapping populations having the phenotypic data of 10 traits were used for the association study.

Pop1

The phenotype and genotype data associated with plant height (PH), dry biomass (DBM), and starch (ST) were adapted from published work (Valluru et al., 2019) for GWAS analysis. A subset of 227 accessions from the 354 WGS set belonged to four major races of sorghum having representation from Africa, Asia, and America was used. The SNPs corresponding to the above-mentioned genotypes were filtered with vcftools and used for GWAS. In 2016, the PH was recorded from 4 to 16 weeks after planting with an interval of 2 weeks, DBM and ST was measured at harvest.

Pop2

The stay-green fine-mapping population developed by crossing an introgression line cross RSG04008-6 × J2614-11 (Usha Kiranmayee et al., 2020) was used for association study using the pan-genome assembly. The DNA from parents and 152 individuals were isolated and skim-sequenced to produce genotype data to a depth of 0.1X. The sequence reads were QC'd with trimmomatic (Bolger et al., 2014), mapped with bowtie2 (Langmead and Salzberg, 2012) and SNP called with GATK (McKenna et al., 2010) and filtered with vcftools (Danecek et al., 2011) as above said method.

The Pop2 was evaluated with green leaf area (GLA) trait in the *rabi* season of 2012–2013 and 2013–2014 at ICRISAT, Patancheru, India. The GLA percentage was measured from seven to 49 days after flowering (DAF) for every 7 days interval in both years. Additionally, in the year 2013, the phenotypes of glossy (GL), leaf sheath pigment (LSP), plant vigour (V), trichome low (TL), trichome up (TU), soot fly dead hearts (SFDH) traits were recorded in rabi (R13), and *kharif* (K13) seasons.

The genotype to phenotype association was performed with GAPIT (Lipka et al., 2012) and the results were initially filtered with Bonferroni cut-off [-log10(p-value)>2.5] followed by p-value and false discovery rate values <0.05 (close to Benjamini-Hochberg cut-off value) as the significant values. These significant SNPs were further functionally annotated with predicted gene coordinates.

Drought RNASeq Assay Analysis

To demonstrate the utility of the pangenome, we have used a sorghum transcriptome experiment on drought response (Abdel-Ghany et al., 2020) available in the Sequence Read Archive (SRP227627). In this study, the RNASeq data were derived from contrasting genotypes- drought resistant [BT × 623 (DR1) & SC56 (DR2)] and drought susceptible [T × 7000 (DS1) and PI482662 (DS2)] at the seedling stage was obtained The quality cheque was performed on raw sequence reads using FastQC (Andrews, 2015) followed by cleaning the low-quality reads and removing sequencing adaptors using the Trimmomatic (Bolger et al., 2014) tool. Trimmed reads were aligned to the Sorghum pan-genome using TopHat2 (Kim et al., 2013) and bam files were filtered to remove reads aligned to multiple locations. Differential gene expression was performed on different conditions using Cuffdiff (Trapnell et al., 2010) to compute logFC and q-values across all accessions at different conditions (control and treated). A total of eight conditions were analysed to find drought-induced genes after 1 and 6 h of post-treatment (20% PEG treatment). Two biological replicates were analysed for each condition resulting in 32 samples (4 genotypes × 2 conditions × 2-time points × 2 replicates). The differentially expressed genes (DEGs) were determined if the $q < 0.05$ and log2FC is <-2 or >2 ratios between control and treatment for each time point and in each genotype.

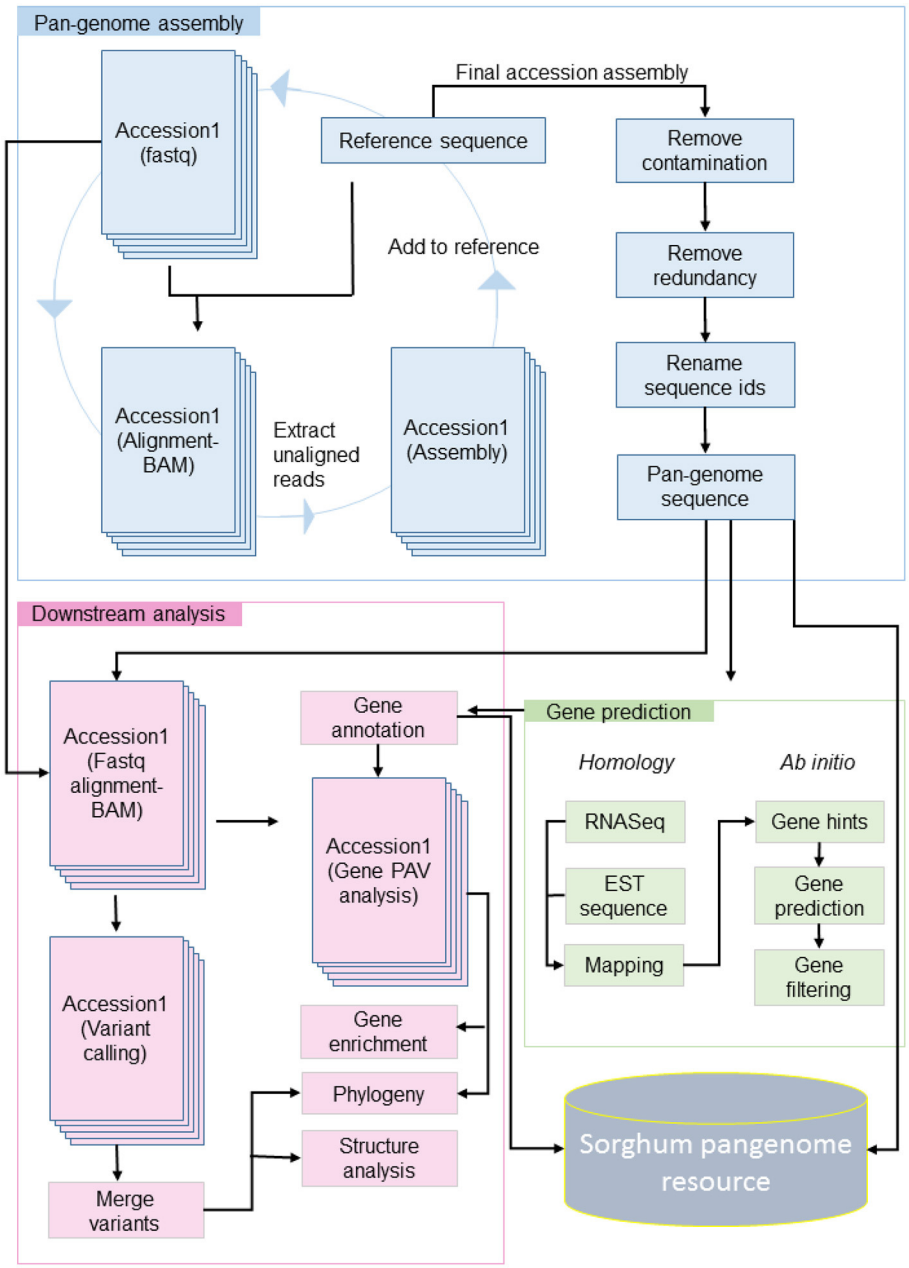

FIGURE 1 | Raw reads of accession were aligned to reference, the unaligned reads were assembled and the contig sequence added to the reference sequence. The updated reference with assembled contigs used as reference for next iteration. The contamination and redundant sequence filtered from final updated reference and genes predicted with homology and *ab initio* method. On availability of pan-genome sequence and gene models, the downstream analysis performed including PAV, variant calling, phylogeny, enrichment, and structure analysis.

RESULTS

Pan-Genome Assembly

Genome sequence data with minimum 10X coverage from earlier studies (Guo et al., 2019; Valluru et al., 2019) were used for pan-genome assembly (**Supplementary Table 1**). The pan-genome was constructed using 176 sorghum accessions using an iterative mapping and assembly approach, similar to Brassica (Golicz et al., 2016a) and pigeon pea (Zhao et al., 2020) (**Figure 1**).

On an average, each iteration of the process added 1.9 Mb of sequence to the reference (**Supplementary Figure 1**) and a total of 263.7 Mbp was assembled. Of these, 89.2 Mb of the sequence were removed as contaminants (including chloroplast and mitochondrial sequences) and/or duplicated contigs. The final resulting pan-genome contained 210,805 contigs with a total length of 883.3 Mb (**Figure 2**) with a minimum contig size of 500 bp. Gene density on the contigs added by this pan-genome exercise was lower than on assembled chromosomes

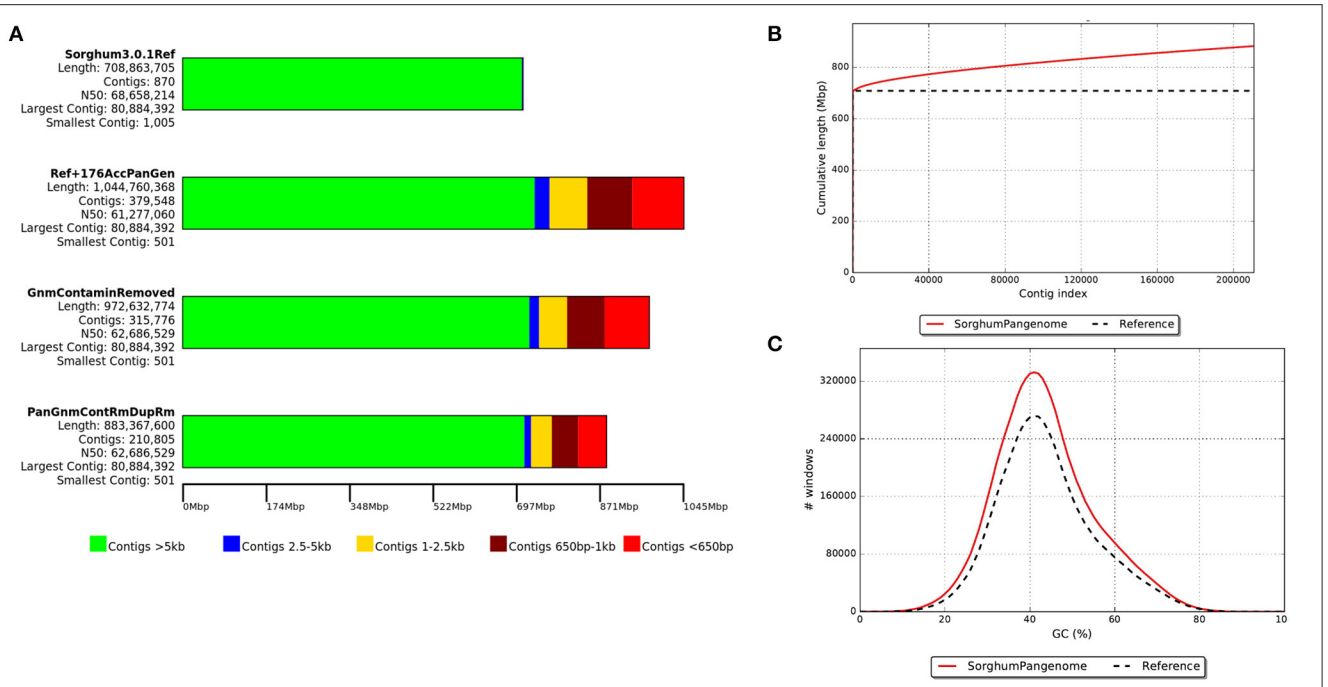

FIGURE 2 | The development of sorghum pan-genome from reference genome assembly: **(A)** a draught pan-genome (including sequence contamination, duplication, chloroplast, and mitochondrial sequence) and final pan-genome assembly sequence size, **(B)** cumulative length of assembled contigs and **(C)** GC percentage of cleaned assembled contig sequences.

but comparable to the density observed on the reference unplaced scaffolds (**Figure 3**).

The pan-genome showed an increase of 24.6% (174.5 Mb) over the reference genome, which was the second-biggest increase of any previously reported pan-genome after the tomato pan-genome. The increase in tomato pan-genome size was captured a 42% non-reference sequence from 725 accessions including the wild relatives (Gao et al., 2019). In other species, an increase in sequence size of 3.3% in wheat (Montenegro et al., 2017), 4% in *Oryza sativa japonica,* 6% in *Oryza sativa indica* (Yao et al., 2015), 5% in *Brachypodium distachyon*, and 20% in *Brassica oleracea* (Golicz et al., 2016a) was documented. The relative increase in sorghum pan-genome assembly size indicated that the presence of high level of genome diversity contributed by the accessions used in this study.

The assembled sequence was annotated using a strategy called combining evidence-based *ab initio* gene prediction. RNASeq (Guo et al., 2019) mapping hints from the 25 accessions used for *ab initio* gene prediction and the 3,589 genes supporting the mapped expressed sequence tags (EST) sequences were retained. We identified 11,057 to 17,616 variable genes in the 176 genomes, with an average gene sequence length and exons per gene of 1,567 bp and 3.6, respectively. The gene length and exons in core genes were more than the variable genes comparatively (**Figure 4**).

Sorghum Pan-Genome Gene PAV (gPAV)

The gPAV in genes among the sorghum accessions could reveal the genetic changes that can be used to infer the phylogenetic history as well as to select the potential targets for breeding. To identify the gPAVs, sequence reads were mapped to the pan-genome contigs and genes were scored as present or absent based on the mapped sequence read coverage (**Supplementary Table 2**, **Figure 5**). For a given gene, to assess the gene loss event, the mapping of the whole genome sequence reads was measured. On an average, each sorghum accession contained 32,795 genes (**Supplementary Table 3**), of which 16,821 (47%) were core genes or in other words, they were shared by all remaining accessions. Comparatively, tomato (Gao et al., 2019) (74.2%), maize (Hirsch et al., 2014) (39%), *Arabidopsis thaliana* (Contreras-Moreira et al., 2017) (70%), wheat (Montenegro et al., 2017) (64%), pigeon pea (Zhao et al., 2020) (86%), *Brassica rapa* (Lin et al., 2014) (87%), *O. sativa* (Schatz et al., 2014) (92%), and, *Brassica napus* (Hurgobin et al., 2018) (62%) had higher number of genes (Bayer et al., 2020). On the other hand, 18,898 genes were variable/accessory genes (**Figure 4A**), of which 30 genes were uniquely present (indicating that the genes are present in any one accession but absent in remaining all accessions) and 3,183 (8.9%) were uniquely absent (indicating that the genes present in all accessions but absent in any one accession) (**Figure 5**). Variable genes were found shorter and had fewer exons per gene when compared to core genes (**Figure 4C**) which were in agreement with *O. sativa* and *A. thaliana* crop studies (Bush et al., 2014; Schatz et al., 2014; Golicz et al., 2016b). Based on gPAVs from 354 cultivars, we estimated the sorghum pan-genome had a closed type of pan-genome (**Figure 4B**), with 30 genes were uniquely present and 3,183 genes were uniquely absent. The uniquely present genes were fewer than

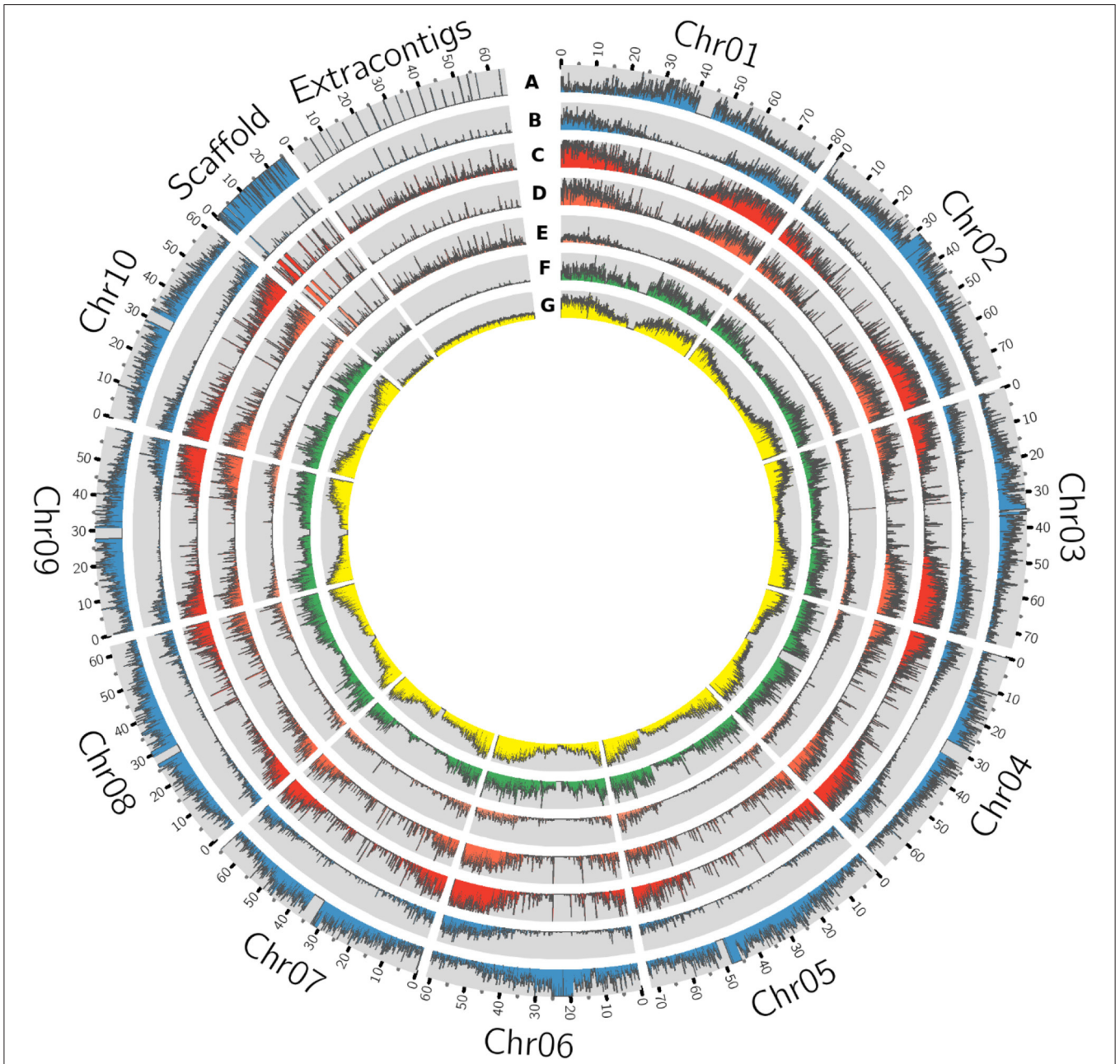

FIGURE 3 | Circos plot of sorghum pan-genome having the extra contigs as a novel sequence assembled from 176 accessions. The genomic features of each track represent absolute values to the respective track (with 80 Kb window size), **(A)** Reference whole genome sequence reads mapping **(B)** Drought expression (RNASeq) sequence mapping density. **(C)** Gene density **(D)** Genes commonly present in all accessions (core genes) **(E)**. genes absent in at least one of the accessions (variable genes) **(F)** SNP density **(G)** Insertions and deletions (indels).

the wheat (49 unique genes per cultivar) (Montenegro et al., 2017) and *B. oleracea* (37 unique genes per cultivar) (Golicz et al., 2016a). Of the 30 genes uniquely present in any single sorghum accession, nine such genes were reported from Macia accession alone (**Figure 5**). Extending the population size and including the wild relatives could further increase the measure of the gene content of this species (**Figure 4B**) (Golicz et al., 2016b).

Gene Functional Analysis

We identified enriched biological pathways by performing gene enrichment analysis using the R topGO package. The significantly enriched pathways related to responsive genes were identified (**Figure 6**). A total of 94 most significantly enriched genes (**Supplementary Table 4**) for biological process pathways are shown in **Supplementary Figure 2**. The gene ontology (GO) enrichment analysis showed that the genes

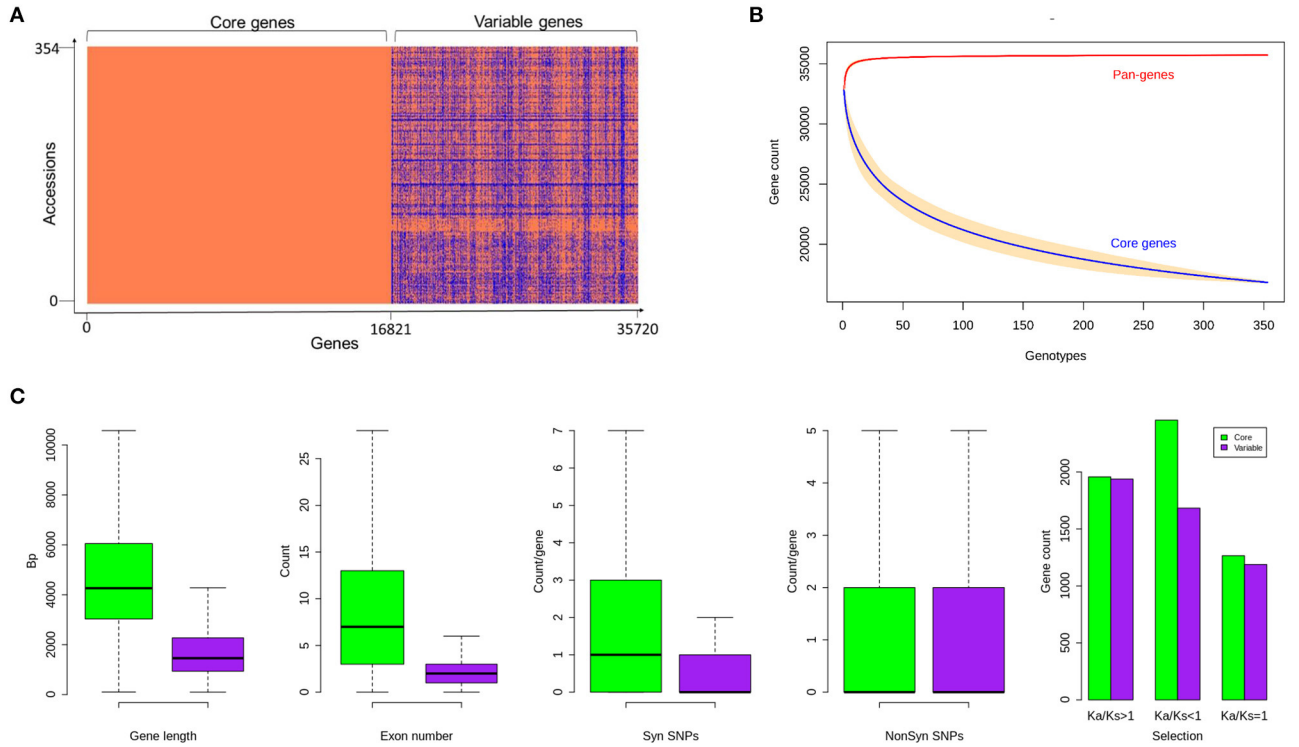

FIGURE 4 | (A) Gene presence and absence variations (gPAVs) in sorghum accessions **(B)** Gene variation of pan-genome and core genome **(C)** Sorghum core and variable gene properties. Variable gene length is shorter with fewer exons. The variable genes have fewer synonymous SNPs and similar non-synonymous SNPs compared to core genes. The gene counts with different Ka/Ks category indicating the selection pressure in both core and variable genes. There is a difference in the number of SNPs between core and variable genes within all groups.

were enriched in response to chemical, hormone, organic substance, stress, auxin and abiotic stimulus (**Figure 6A**). It was noted that most of the pathways were related to plant response to stimulus and chemicals. The gene enrichment among stress responses genes including water deprivation (GO:0009414), desiccation (GO:0009269), abiotic stimulus (GO:0009628), chemical stimulus (GO:0042221), and stress (GO:0006950) were reported in a reference set of genes (Woldesemayat and Ntwasa, 2018). The gPAV-based enrichment on assembled genes from the sorghum pan-genome has added the response of the genes to auxin (GO:0009733), hormone (GO:0009725), organic substance (GO:0010033), hypoxia (GO:0001666), and decreased oxygen levels (GO:0036294). The functional annotation of the variable genes highlighted the genes involved in response to biotic and abiotic stress indicating the possible evolutionary adaptive traits (Lasky et al., 2015). Macia (9 genes), Ajabsido (4 genes), and PI329719 (4 genes) were identified with a number of trait-specific genes (**Figure 5**), which could be used as potential donors for trait improvement. It was observed that the above-mentioned unique genes were involved in response to the stimulus (GO:0050896), chemical (g8132, GO:0042221), and arsenic-containing substance (g24192, GO:0046685) (**Supplementary Table 5**).

Variant Discovery

We identified a large number of variants (single nucleotide polymorphism (SNPs) and indels) by mapping the sorghum population whole-genome sequence reads to sorghum pan-genome assembly using GATK. Of the total of 2.0 million SNPs, 91,319 were in the extra contig (assembly) sequence (**Supplementary Table 6**). The SNP density in extra contigs (0.52/Kbp) was much less compared to the density in the reference genome assembly (2.72/Kbp) (**Figure 3**). The SNP annotation results illustrated the highest number of SNPs in intergenic region (40%) followed by upstream (22.5%), downstream (21.4%), intron (8.8%), and exon (3.6%) regions with an overall Ts/Tv ratio of 1.92. Chromosome 4 had the highest number of SNPs (251,830), followed by chromosomes 1, 2, 3, 5, 10, 8, 6, 9, and 7, respectively. Chromosome 7 had the fewest number of SNPs (119,019) with the highest density of 0.55/Kbp and chromosome 4 had the least density of 0.27/Kbp (**Supplementary Table 6**) (**Figure 3**). The presence of more SNPs and indels in the telomeres compared to centromeres explained the higher gene activity toward the telomeres supporting the SNPs and indels density reported in historically important grain sorghum genotypes (BTx623, BTx642, and Tx7000) (Evans et al., 2013) (**Figure 3**). The SNP annotation reported the frequency of synonymous SNPs in the core genes was much higher than

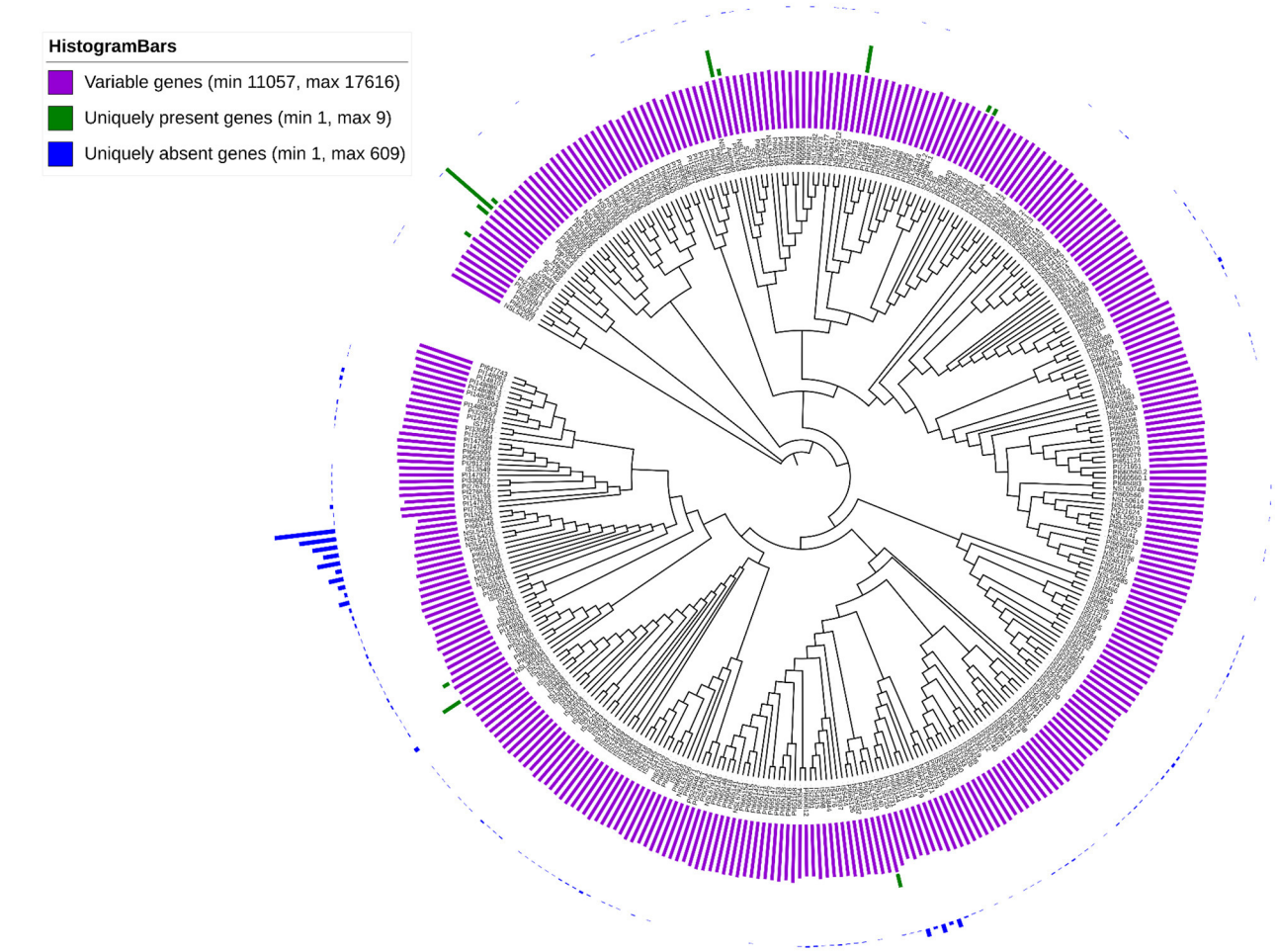

FIGURE 5 | gPAVs-based neighbour-joining tree shows the genetic relationship among the sorghum accessions. Histogram shows the variable genes (purple bars), uniquely present genes (green bars) and uniquely absent genes (blue bars).

in the variable genes (**Figure 2**). This was in contrast to the higher mis-sense SNPs in core pigeon pea genes to variable genes early reported (Zhao et al., 2020). We detected genome-wide indels of various size (**Supplementary Figure 3A**) and the genes featuring indels has reduced proportionally to the size of the indels (**Supplementary Figure 3B**). On increasing the indel size, the number of the indels decreases in both gene and genome-wide sequence. The overall indels count from the sorghum accessions used in this study was much higher than the indels earlier reported in the six sorghum accessions (Yan et al., 2018). A total of 36,097 genes had 983,060 CNVs among the sorghum accessions used in this study. The Ka/Ks ratio estimating the balance between neutral mutations, purifying selection, and beneficial mutations on a set of core and variable genes exhibited that, core gene count under positive selection were significantly close to variable gene count compared to genes under the negative selection pressure (**Figure 4**).

The maximum (432,286) and minimum (2,854) number of SNPs were identified in sorghum accessions PI267614- NSL54318 and IS3693- IS23514, respectively (**Supplementary Table 7**). The accessions NSL54318 (849,052) and IS3693 (17,084) had the maximum and minimum polymorphism, respectively, with sorghum pan-genome assembly sequence (**Supplementary Table 8**). The SNPs were validated with 3K SNPs Infinium array (Bekele et al., 2013) and of the 2,980 mapped flanking SNPs sequence, only 20 did not map on pan-genome assembly. The overall alignment rate was 99.33%, from the mapped 2,980 SNPs array sequences (**Supplementary Table 9**, **Figure 6B**). Among them, 37 SNP sequences were mapped to extra contigs (novel sequence assembly) and 150 (5%) did not represent any GATK SNPs calls (29 SNPs from extra-contigs). In addition to the core SNPs of the array sequence, more SNPs were identified in the flanking sequence. Out of 15,383 GATK SNPs on the mapped array sequence, 15,314 SNPs were validated with the GATK called

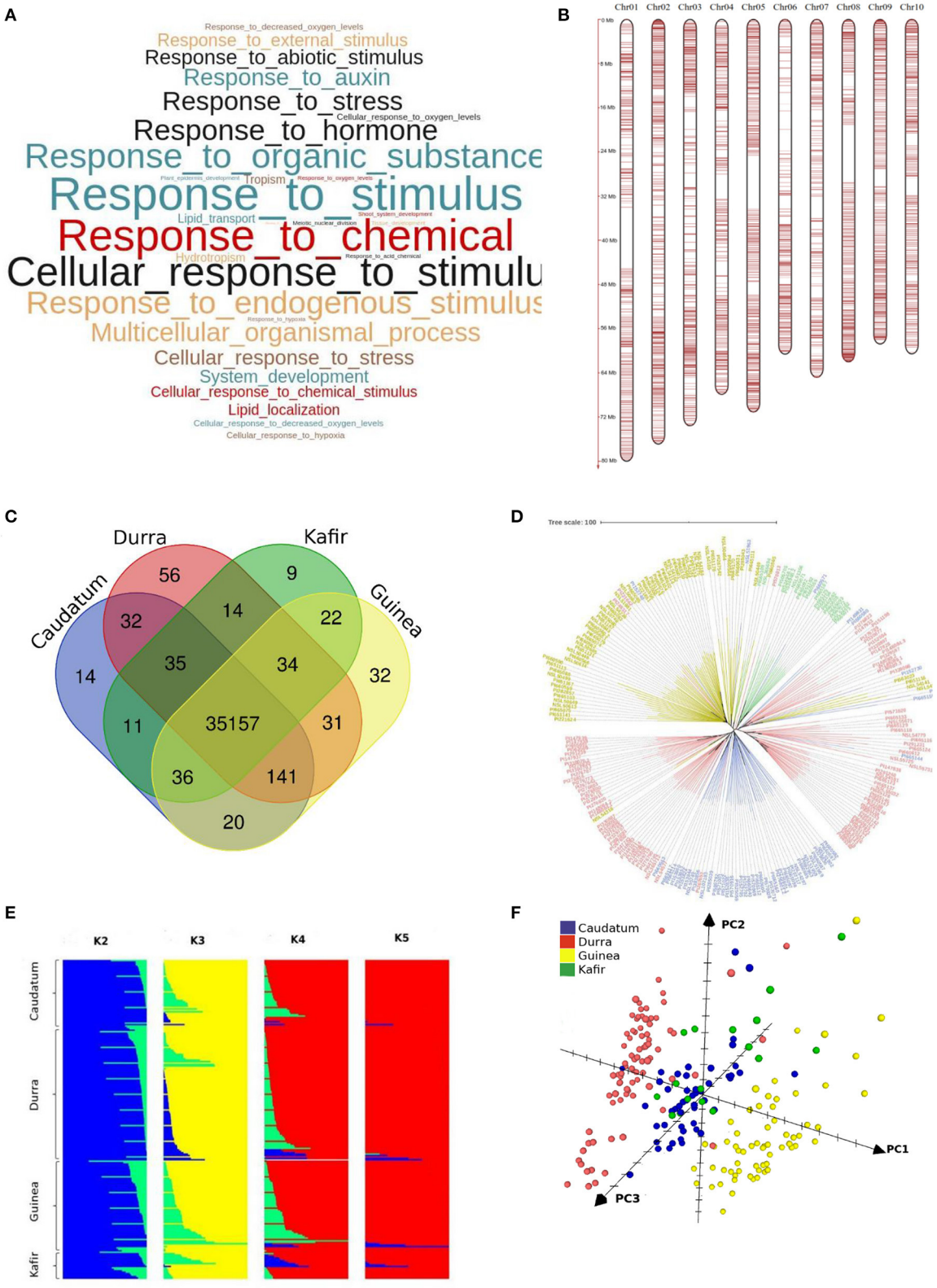

FIGURE 6 | (A) Significantly enriched top GO terms among the variable genes ($p < 0.05$) **(B)** distribution of Infinium SNP array markers on chromosomes **(C)** specific and common genes across races **(D)** neighbour-joining tree shows the genetic relation among the sorghum accessions belonged to different races (blue-*Caudatum*, red-*Durra*, green-*Kafir* and yellow-*Guinea*) **(E)** structure analysis of sorghum population with K2 to K5 and **(F)** sorghum accessions grouped by *caudatum*, *durra*, *guinea* and *kafir* race through principal co-ordinate analysis (PCo).

allele (**Supplementary Table 9**). Finally, on validation with array SNPs, the overall GATK SNP calling reported 99.9% accuracy.

To understand the genetic relationship of the 354 sorghum accessions, a neighbour joining (NJ) tree was constructed with the SNPs (**Supplementary Figure 4**). The accessions were arranged in many sub-groups indicating the possible sorghum race accessions. To assess the race-specific accessions, the 216 known sorghum race accession bootstrapped to construct an NJ tree. The NJ tree showed the subgroups of sorghum accessions according to the races except a few, indicating the hybridisation process in the past (**Figure 6D**). For example, PI221662, a durra race accession was genetically related to the *guinea* race. Similarly, to understand the gene PAV-based genetic relation, the phylogenetic relationship among sorghum accessions was assessed by distance-based 1,000 bootstrap replicates and represented through the NJ tree. Among the 35,719 total genes, 53% exhibited the genic variations to estimate the relationship among the accessions (**Figure 5**). The largest number of genes uniquely present and absent genes was found in Macia (9 genes) and PI660645 (372 genes), which indicated the evolutionary distance from other accessions.

With the known four races (Obilana et al., 1996), the structure analysis with the variants set showed the presence of three sub-population (**Figure 6E**), resulting an expected admixture between *caudataum* and *kafir* accessions which was in agreement with early study (Valluru et al., 2019) (**Figure 4**). The result was also validated by the PCo, where *durra* and *guinea* sorghum races displayed identifiable clusters, because of the available sequence representation through pan-genome, while *caudatum* and *kafir* accessions exhibited the admixers (**Figure 6F**). The earlier principal component study shows the mixed grouping of *guinea* and *kafir* accession in the (Sapkota et al., 2020), indicating the missing sequence representation for all race in the single reference genome.

Variation of Sorghum Race-Variable Genome

Sorghum pan-genome analysis has identified 18,898 variable genes, and the gene cluster analysis identified 11,470 gene families, of which un-clustered genes (6,057) included 556 from the non-reference genes and the remaining 5,501 were reference genes. Among these un-clustered genes, 3,195 were orthologous to *Zea mays, Setaria italica, Brachypodium distachyon,* and *Oryza sativa* and the remaining 2,862 were paralogous. Among the total variable genes, a total of 111 genes are race-specific and the gene shares among four sorghum races showed that the *durra* and *guinea* had a maximum of 56 and 32 unique genes, respectively, making them more diverse than the other two races with 14 (*caudatum*) and 9 (*kafir*) unique genes (**Figure 6C**). The gene annotations suggested that the unique genes from *durra* were associated with heat shock protein, LRR repeat protein, L-type lactin-domain receptor, ABC transporter family proteins, and Ras-related proteins. *Guinea* group had the unique genes associated with disease resistance protein, beta-glucosidase proteins, NRT1/PTE protein family, and Alpha/beta-Hydrolases superfamily proteins (**Supplementary Table 10**). The gene uniqueness to specific races possibly reflected the selection of the genotypes for adapting to the respective ecological conditions (Upadhyaya et al., 2017).

GWAS

Two populations namely, Pop1 (Valluru et al., 2019) and Pop2 (Usha Kiranmayee et al., 2020) were used for GWAS to understand the functional utility of the pan-genome. Pop1 had 216 accessions with the phenotypes of DBM, PH, and ST while Pop2, a stay-green fine-mapping population with 190 segregates, had green leaf area (GLA), GL, V, LSP, SFDH, TL, and TU.

In Pop1, the SNPs were further filtered by accessions and on applying the SNP quality philtres, which retained 1.12 million SNPs for association analysis. Pop2 having sequence data of 190 genotypes processed to map to pan-genome and 109,338 SNPs were used for GWAS.

We identified a total of 397 unique SNPs having significant association (having *p*-value and false discovery rate below 0.05) in both Pop1 and Pop2 traits, of which 216 SNPs were commonly mapped with multiple traits. Most of these SNPs distributed on chromosome 10 (120 SNPs) followed by chromosome 6 (69 SNPs). The reference genome alone had 385 SNPs and the rest of the SNP-trait associations located on the unmapped read sequence assembly.

For the Pop1, a total of 36 SNPs had a significant association across three traits (**Figure 7A**) (**Supplementary Table 11**). Among them, seven were located on newly assembled contigs (DBM and PH) (**Figure 7B**), three were from unplaced reference contigs and the remaining 26 are from chromosome sequence. Among the 36 linked SNPs, 10 were genic and the remaining 26 were inter-genic regions (**Supplementary Table 12**). Three of the genic SNPs were associated with DBM while six were associated with PH and the remaining one co-mapped to both DBM and PH. From the 10 associated genes, three genes (Sobic.002G022500, Sobic.003G173400, and Sobic.004G350800) were from the core gene set and the remaining belonged to variable genes.

From the Pop2, the GLA, at various stages (**Supplementary Table 13**), associated with 219 SNPs including 111 genic, distributed across all chromosomes including the pan-genome assembly contigs (**Supplementary Table 11**). GL, LSP, SFDH, TL, and TU traits were associated with 129 and 103 significant SNPs in *Rabi* (R13) and *Kharif* (K13) seasons, respectively (**Supplementary Table 11**). The majority of the SNPs were associated with chromosome 10 followed by 5 and 6 in both the seasons. Among them, a total of 96 SNPs was mapped across seasons and a total of 18 and 196 showed season-specific association in K13 and R13, respectively. Interestingly, only four genic SNPs were associated with TU in K13, whereas 63 were associated in R13 explained the season-specific gene regulations. Similarly, SFDH had no association in K13 but had 56 genic SNPs in R13 season (**Supplementary Table 11**).

The number of SNPs associated with DBM, PH, ST (Pop1), plant vigour (V), GL, LSP, SFDH, TL and TU, and GLA (Pop2) was 10, 25, 1, 1, 23, 31, 84, 169, 98, and 397, respectively. Among the chromosomes, as many as 392 of the SNPs were

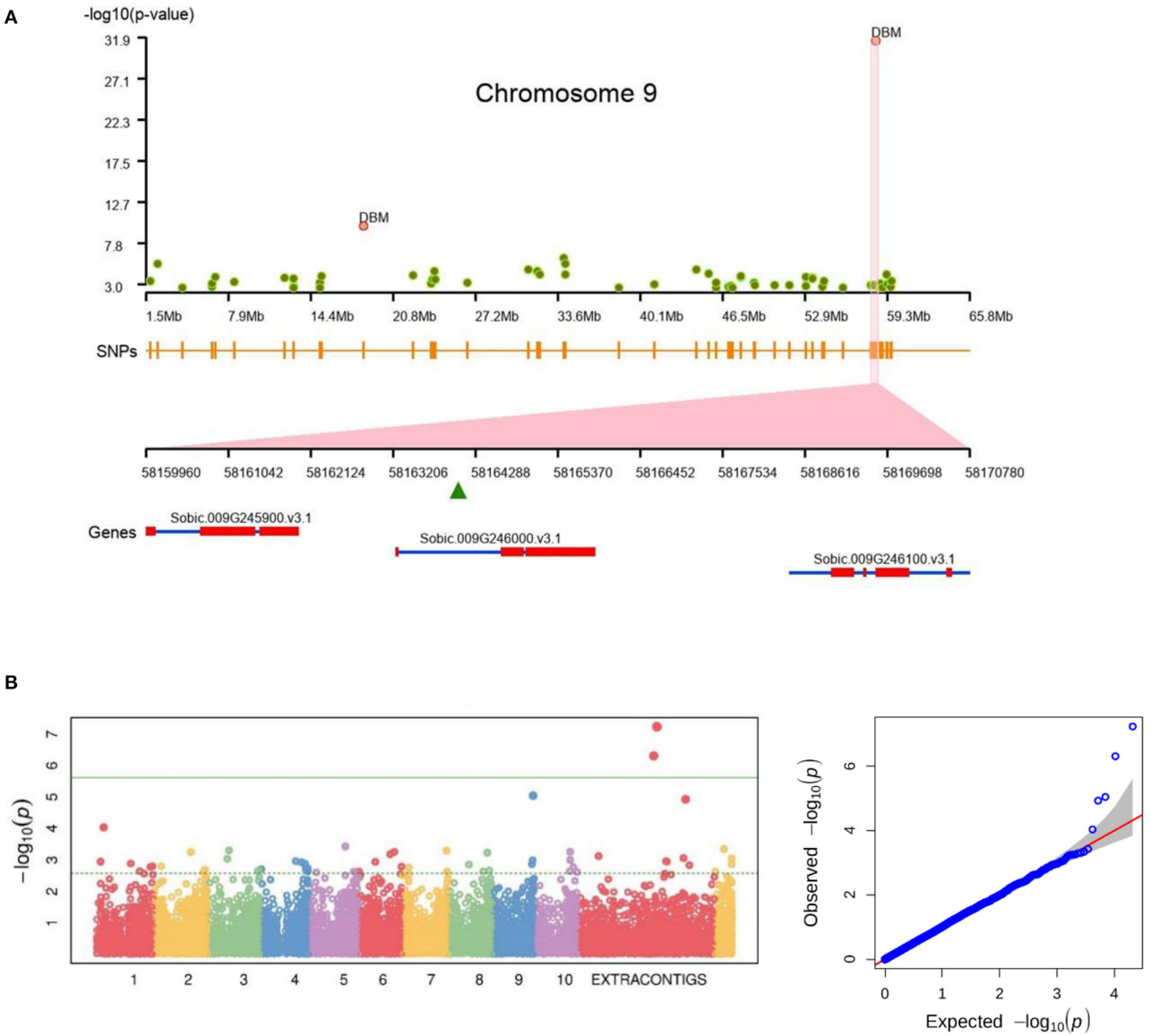

FIGURE 7 | Genome-wide association showing the significant association of SNPs for **(A)** plant biomass on chromosome 9 (on Sorbic.009G246000.v3.1) and **(B)** plant height trait association on extra-contigs.

associated with chromosome 10 and only 8 SNPs were associated with chromosome 3 (3 SNPs on scaffolds). The pan-genome assembly contigs hold 15 trait-associated SNPs, an additional genetic resource for the sorghum breeding program.

Of the total 183 GWAS SNPs directly associated with gene functions, the DBM and PH (from Pop1) were associated with 10 genes (off these, 1 gene assembled in this study). In Pop2, 173 genes were distributed as 96, 11, 13, 46, 48 and 1 for GLA, GL, LSP, SFDH, TL, and TU, respectively.

Identification of the Drought Candidate Genes

A sorghum RNASeq data generated from drought-resistant [BTx623 (DR1) & SC56 (DR2)] and susceptible [Tx7000 (DS1) and PI482662 (DS2)] genotypes at different seeding stages (Abdel-Ghany et al., 2020) were re-analysed and mapped through the newly developed pan-genome. A total of 1,788 genes were significantly affected by drought stress (**Figure 3**) and among them, 79 genes were reported from genes on assembly sequence (extra- contig) (**Figure 8A**) (**Supplementary Table 14**).

The drought-resistance (DR1 and DR2) and drought-susceptibility (DS1 & DS2) samples were phenotyped at two conditions (1 and 6 h). The DR1 and DR2 samples reported (1 h treatment) a total of 590 (450 up and 140 down-regulated) and 195 (180 up- and 15 down-regulated) expressed genes, respectively (**Figure 8B**). Of these, none of the genes reported from the novel sequences, indicating both (DR1 and DR2) were closely related. Additionally, DR1 and the reference sequence

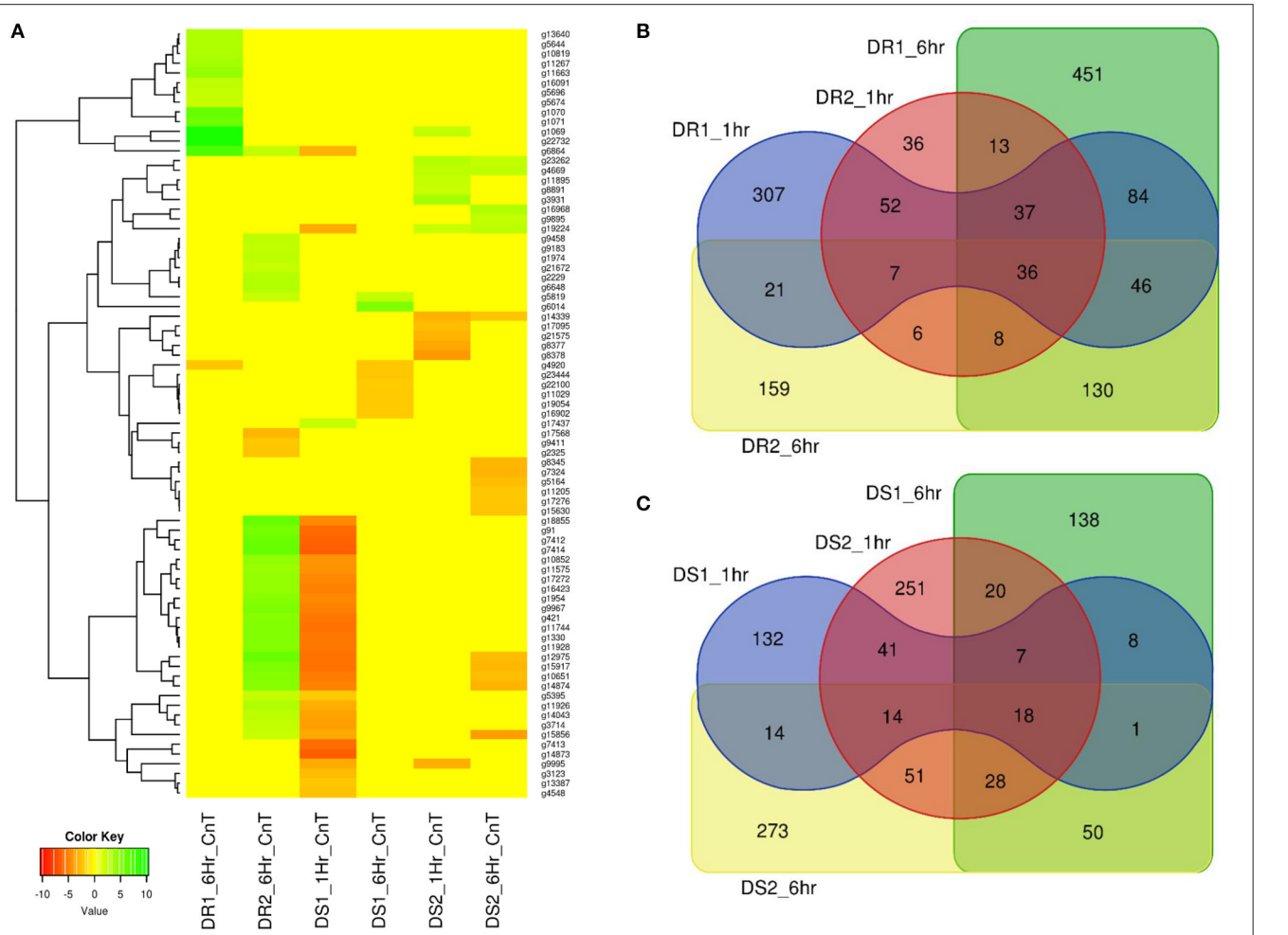

FIGURE 8 | (A) Sorghum drought trait-based RNASeq assay on pan-genome assembled genes (extra contigs). DR1 and DR2 were the two datasets for drought resistance after 1 and 6h, respectively. DS1 and DS2 related to drought susceptibility after 1 and 6 h, respectively. Heatmap showing the range of up-regulated genes as green and down-regulated genes as red colour **(B)** Drought resistance in BTx623 (DR1) and SC56 (DR2) and **(C)** Tx7000 (DS1) and PI482662 (DS2) drought susceptible genes share between the genotype and treated conditions (1 and 6 h).

belong to the same genotype and this supports the absence of gene expression from the novel sequence at this condition. When the treatment was extended for 6 h, 14 (13 up and 1 down-regulated) and 34 (31 up and 3 down-regulated) genes from novel sequence were expressed for DR1 and DR2 data-sets, respectively.

Similarly, DS1 and DS2 samples showed (for 1 h treatment) 235 (123 up and 112 down-regulated) and 430 (388 up and 42 down-regulated) genes, respectively (**Figure 8C**). Of these, DS1 and DS2 had 32 (1 up and 31 down-regulated) and 13 (7 up and 6 down-regulated) expressed genes from novel sequence, respectively. After 6 h of treatment, DS1 and DS2 samples reported 270 and 449 expressed genes respectively. Of these, DS1 and DS2 reported 8 (2 up and 6 down-regulated) and 17 (5 up and 12 down-regulated) expressed genes were from the novel sequence, respectively.

Over-all, five drought-related genes were co-mapped with the trait-associated genes. Among the five genes, three traits-linked genes Sobic.001G363200 (GLA), Sobic.007G180300 (GL), and Sobic.010G231900 (TL, TU, and SFDH) were commonly expressed in drought resistance and susceptibility conditions. The remaining two drought resistance specific genes Sobic.005G069800 and Sobic.006G127800 were linked to PH and LSP traits.

DISCUSSION

We built a sorghum pan-genome with an iterative mapping and assembly approaches with 176 of 354 whole genomes sequenced accessions having coverage of more than 10X. The total size of the pan-genome has become 883 Mbp, with a 20% increase (175 Mbp) compare to the reference assembly of 708 Mbp. This level of novel sequence increase probably due to the high level of genetic diversity observed in the respective species (Cuevas and Prom, 2020).

We have generated the pan-genome genomic resource from the diverse sorghum accessions including the basic

and intermediate sorghum races [bicolor (B), *caudatum* (C), *durra* (D), *kafir* (K), and *guinea* (G)]. Comparison of the wide range of sorghum whole genome sequence datasets has enabled to assemble many coding genes that were absent in published sorghum reference genome sequences. The mapping of RNASeq read from 25 accessions on the assembled contigs supports the predicted genes on the novel sequence (**Supplementary Figure 5**) is an additional genetic resource that will enhance the identification of the quantitative trait locus (QTL) and genome-wide association studies (Chen et al., 2014; Yano et al., 2016; Zhao et al., 2020). The earlier pan-genome studies found that non-reference genes have significantly involved in agronomic traits mainly in plant defence responses (Hirsch et al., 2014; Golicz et al., 2016a; Montenegro et al., 2017; Dolatabadian et al., 2020). Similar to the sorghum genes, *B. oleracea* pan-genome genes also showed that nearly 30 percent of reference genes exhibited the gPAV (Golicz et al., 2016a). It is understood that, as the number of genotypes increases, the size of core genes decreases with a relative increase of variable genes (**Figure 4B**). With the 10 sample sizes, *Brassica oleracea* pan-genome had 20% of PAV genes (Golicz et al., 2016a) which was in consistent with simulation with similar population size in *B. distachyon* pan-genome (Gordon et al., 2017). Similarly, pan-genome from 15 Medicago genomes had 42% of sequences share with few accessions (Zhou et al., 2017), which was comparatively similar to the 49% of sorghum variable pan-genes in this study.

The result of the structure groupings correlated with the PCo showing three different clusters with one of them having two groups (*caudatum* and *kafir*). Among the four basic sorghum races used in this study, PCo displayed, *guinea*, and *durra* remain as distinct clusters while *caudatum* with *kafir* classified with mixed genotypes, which is considered as the stable hybrid race among the 10 possible stable combinations of sorghum races (Obilana et al., 1996). Similarly, mixed PCo clusters were also reported earlier with five basic sorghum races, where the sorghum B race was not well-supported genetically and a majority of them share membership with the remaining four genetic groups (Brown et al., 2011).

The genomic features helped the races to group into different clusters. By looking at the race-specific genomic data, each race had distinctive features. The *guinea* group had 37 accessions with race specific genes present in range of 2–13 genes per accessions, whereas *durra* had 2–12 genes in 92 accessions, *kafir* had 2–5 genes in 12 accession, and *caudatum* had 2–4 genes in 15 accessions. The two groups including *durra* and *guinea* were having 56 and 32 distinct genes, respectively, unique to these groups, whereas *caudatum* and *kafir* have on 14 and 9 distinct unique genes, as these groups have the admixture accessions which share genes between the groups.

The functional analysis of variable genes was enriched with GO terms associated with response to light, flower development, salt stress, water, heat, desiccation, temperature, osmotic stress, lipid, gibberellin, and stress. The results supports the earlier gene function based clustering and enrichment analysis exhibiting the similar stress response genes reported in sorghum (Woldesemayat and Ntwasa, 2018). The plant hypersensitive response annotation in the variable gene was reported in plant pan-genome analysis (Golicz et al., 2016a; Montenegro et al., 2017; Hurgobin et al., 2018; Zhao et al., 2020).

The development and application of sorghum SNPs have limited to reference genome assembly sequence used in the analysis. The 1.8 million SNP reported earlier on Rio with respect to BTx623 (Cooper et al., 2019), were limited to the single reference genome. Using the whole genome sequence data from 354 sorghum diverse accessions, we identified two million SNPs and 3.9 million indel sites, which represented the functional genome diversity. The density of genetic variation in the novel assembled sequence was low compared to the reference sequence. The reference genome carried most of the conserved essential genes, indicating that the variable sequence has low diversity (**Figure 3**), as reported in the six sorghum accessions from common geographical regions (Yan et al., 2018). The fewer number of SNPs on variable sequence mainly contained genes involved in response to various stress (biotic and abiotic stress tolerance), this finding is well-aligned with the SNPs from disease resistance R genes differentiating sweet and grain sorghum accessions (Zheng et al., 2011). A reference sequence within the pan-genome assembly alone accounted for 95.4% of SNPs and the added assembly sequence from the sorghum population had 4.5% additional SNPs. A total of 2,980 array SNPs from (Bekele et al., 2013) were identified as similar to GATK called (reference-based variant calling) SNPs with 99.33% of true SNPs. The GATK called sorghum SNPs validation rate with array SNPs was higher (99.33%) compared to the non-reference based variant calling methods, for example, the wheat pan-genome SNPs were called with 96.3% accuracy (Montenegro et al., 2017). The abundance of SNPs depends on factors such as mutation events and genome diversity and the SNPs identified in the variable genome can assist in characterising novel metabolic pathways.

Phylogenetic analysis of 354 sorghum accessions using SNPs on the pan-genome demonstrated the mixed groups of diverse biomass genotypes (Valluru et al., 2019), domesticated accessions (Guo et al., 2019) and Chibas sorghum breeding program accessions (Jensen et al., 2020). gPAV-based phylogeny showed a group of 15 accessions having uniquely absent genes in a range of 2–509 genes from the biomass genotypes indicating the wider genetic diversity. The five Chibas sorghum breeding lines (Macia, Ajabsido, SC1345, P898012, and Grassl) had the most unique genes followed by seven domesticated accessions distributed across the phylogenetic tree. On assessing the known sorghum race genotypes from Valluru et al. (2019) phylogeny showed a cluster for each sorghum race. Few accessions of *caudatum* and *guinea* were mixed with *durra* cluster indicating that these are the *caudatum-durra* (CD) and *guinea-durra* (GD) hybrid individuals. Similarly, few accessions were not placed in respective race groups, for example, PI248317 accession was a *durra* race accession placed in *guinea* race cluster which shared the genetic similarity with *guinea* race as DG hybrid individual.

The GWAS performed in the earlier study was limited to the phenotype association only with limited SNPs on the reference genome used (Morris et al., 2013; Kimani et al., 2020). The SNP calling on sorghum pan-genome has enabled the identification of the variants also from non-reference sequence assembly from the genetically diverse accessions. A total of 91,339 SNPs reported

from the assembled sequence were the additional markers used for GWAS. A total of 36 SNPs (from Pop1) were associated with target traits. Among them, 10, 25 and 1 were from assembly sequence (extra contigs) were associated with DBM, PH, and ST, respectively. Additionally, the GLA (from Pop2) had a significant association with five SNPs on extra-contigs. The GLA phenotypes in 2013 and 2014 after 7, 14, 21, 28, 35, 42, 49 days after flowering (DAF) in *rabi* were associated with 219 SNPs. Most of the SNPs were linked with the GLA recorded at the early stage of 7 (linked with 150 SNPs) to 14 DAF (linked with 161 SNPs) (**Supplementary Table 11**). From the flowering stage to 14 days of post-flowering, the GLA expression was significantly linked with 101 common SNPs (two SNPs reported from Extra-Contig101123 at 855 positions and Extra-Contig170379 at 501 base position) (**Supplementary Table 11**). For the phenotypes of GL, LSP, SFDH, TL, and TU in both *rabi* and *kharif* seasons, a total of 147 SNPs were identified, of which 85 were co-mapped in both seasons, 44 were unique to *rabi* and 18 were unique to *kharif*. Out of total 397 associated SNPs, 12 SNPs from novel sequences having significant trait association is an additional gain from the pan-genome assembly.

Most of the associated SNPs linked to genes including NAC-domain protein controls the flowering time and stress response, BTB domain for protein-protein interaction, PSII protein complex for oxygenic photosynthesis, AAI domain protein for lipid transfer protein (LTP). The genes are transcription factors (TFs) such as nuclear TF, reverse transcriptase Ty1/Copia-type domain and BZIP. The genes also associated with ubiquitination pathway proteins such as B-box, F-box, U-box, RING-type, and RING-type E3 ubiquitin transferase protein supporting the sorghumFDB gene family classifications (Tian et al., 2016).

We found 1,788 drought-responsive genes with different seeding stage sequence data mapping on pan-genome assembly, whereas weekly sampled the growing plants and mapping the RNASeq data to reference alone reported the 44% of genes exhibiting the response to drought stress (Varoquaux et al., 2019).

This difference in drought expression was expected between the seedling samples (in 1–6 h difference) compared to root and leaf large scale sampling in 2–17 weeks of pre and post-flowering drought responses (Varoquaux et al., 2019). Similar drought stress gene expression changes were seen in laboratory and greenhouse studies in sorghum genotypes (Johnson et al., 2014; Fracasso et al., 2016). Identifying 79 drought-linked differentially expressed genes on assembly sequence are the additional genes added from this study (**Supplementary Table 15**). These additional genes through pan-genome were mainly involved in the cell membrane, catalytic activity, molecular function regulation, response to the stimulus, metabolic process, cellular, and biological regulation.

The sorghum pan-genome assembly, genes with its annotations, SNPs data sets are available at the sorghum website (https://doi.org/10.21421/D2/RIO2QM).

CONCLUSIONS

We constructed and characterised the sorghum pan-genome using the reference genome assembly and the whole-genome sequence reads of genetically diverse sorghum accessions. The pan-genome had 35,719 predicted genes, which were categorised as core, conserved genes, and variable genes as they exhibited presence and absence variation. The variable genes were enriched with genes response to various stresses. The SNP Infinium array result showed 99% of representation on the pan-genome assembly sequence. About two million SNPs were developed through pan-genome which can use for functional downstream research. The pan-genome resources were validated by assessing the genetic diversity of sorghum races, identification of genes from GWAS and RNASeq studies. These newly generated genomic resources could be used in sorghum genetic gain improvement programs.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s. The sorghum pan-genome assembly and annotation are available at dataverse.icrisat.org (https://doi.org/10.21421/D2/RIO2QM).

AUTHOR CONTRIBUTIONS

PR and AR conceived and designed the project. PR, PG, and SS carried out the analysis. SS managed computational resources and data management. SD provided sorghum trait data. PR, NT, DE, MG, BN, RG, and AR jointly wrote the manuscript. BN, EM, RD, DO, and HG reviewed the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

ACKNOWLEDGMENTS

The authors thank Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, New York, USA and Plant Genome Mapping Laboratory, University of Georgia, Athens, Georgia, USA for providing the sorghum WGRS, RNASeq data, and phenotype data. The authors also acknowledge the supporting funds from AVISA (OPP1198373) and ICAR-BMGF (101165).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.666342/full#supplementary-material

Supplementary Figure 1 | Whole genome sequence of 176 sorghum accessions mapped iteratively to the updated reference sequence assembly and the unmapped sequence reads were assembly iteratively. The plot represents the size of the sequence assembly gained from respective accessions.

Supplementary Figure 2 | The sorghum pangenome variable genes enrichment analysis and the metabolic pathways for **(A)** Biological process **(B)** Molecular function and **(C)** Cellular components. Top 10 GO terms identified for scoring GO terms for enrichment. Rectangles indicate the 10 most significant terms and the colour represents the relative significance ranging from red (most significant) to yellow (least significant). Each node has GO identifier and name with raw p-values and number of significant genes out of total genes annotated.

Supplementary Figure 3 | Insertions and deletions of various size distribution at sorghum pangenome **(A)** genome level and **(B)** gene level.

Supplementary Figure 4 | The genetic relationship of 354 sorghum accession neighbour joining (with 1000 bootstrap) analysis (the unrooted tree with branch length in red colour).

Supplementary Figure 5 | Sorghum 25 accessions (names and NCBI accessions) RNASeq read mapping density on pan-genome assembly for **(A–J)** Chromosome 1-10 **(K)** Scaffold sequence put together as single sequence and **(L)** Non reference sequence assembly contigs from sorghum accessions concatenated to single sequence as extra contig sequence.

Supplementary Table 1 | Raw sequence data used for sorghum pangenome and SNP calling analysis.

Supplementary Table 2 | The summary of gene presents in each sorghum accession.

Supplementary Table 3 | Gene count per each sorghum accession.

Supplementary Table 4 | Gene enrichment based on GO terms among the variable genes.

Supplementary Table 5 | Significantly enriched genes for BP, MF and CC components.

Supplementary Table 6 | Number of SNPs and density per each chromosome sequence.

Supplementary Table 7 | SNP count for pair-wise combination of sorghum accession.

Supplementary Table 8 | SNP count per sorghum accession.

Supplementary Table 9 | SNP validation with sorghum 3k SNP array.

Supplementary Table 10 | Sorghum genes unique to caudatum, durra, kafir, guniea.

Supplementary Table 11 | Genome wide association analysis of the SNPs showing the significant association with traits.

Supplementary Table 12 | Significantly associated SNPs from the genic region.

Supplementary Table 13 | Sorghum green leaf area trait data collection description.

Supplementary Table 14 | Drought RNASeq assay correspondence on sorghum pangenome extracontigs.

Supplementary Table 15 | Drought response genes count on pangenome assembly.

REFERENCES

Abdel-Ghany, S. E., Ullah, F., Ben-Hur, A., and Reddy, A. S. N. (2020). Transcriptome analysis of drought-resistant and drought-sensitive sorghum (Sorghum bicolor) genotypes in response to PEG-induced drought stress. *Int. J. Mol. Sci.* 21:772. doi: 10.3390/ijms21030772

Alexa, A., Rahnenführer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 22, 1600–1607. doi: 10.1093/bioinformatics/btl140

Andrews, S. (2015). FASTQC a quality control tool for high throughput sequence data. *Babraham Inst*. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., and Edwards, D. (2020). Plant pan-genomes are the new reference. *Nat. Plants* 6, 914–920. doi: 10.1038/s41477-020-0733-0

Bayer, P. E., Hurgobin, B., Golicz, A. A., Chan, C. K. K., Yuan, Y., Lee, H. T., et al. (2017). Assembly and comparison of two closely related Brassica napus genomes. *Plant Biotechnol. J.* 15, 1602–1610. doi: 10.1111/pbi.12742

Bekele, W. A., Wieckhorst, S., Friedt, W., and Snowdon, R. J. (2013). High-throughput genomics in sorghum: from whole-genome resequencing to a SNP screening array. *Plant Biotechnol. J.* 11, 1112–1125. doi: 10.1111/pbi.12106

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Brown, P. J., Myles, S., and Kresovich, S. (2011). Genetic support for phenotype-based racial classifi cation in sorghum. *Crop Sci.* 51, 224–230. doi: 10.2135/cropsci2010.03.0179

Bush, S. J., Castillo-Morales, A., Tovar-Corona, J. M., Chen, L., Kover, P. X., and Urrutia, A. O. (2014). Presence-absence variation in A. thaliana is primarily associated with genomic signatures consistent with relaxed selective constraints. *Mol. Biol. Evol.* 31, 59–69. doi: 10.1093/molbev/mst166

Chen, W., Gao, Y., Xie, W., Gong, L., Lu, K., Wang, W., et al. (2014). Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nat. Genet.* 46, 714–721. doi: 10.1038/ng.3007

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* 6, 80–92. doi: 10.4161/fly.19695

Contreras-Moreira, B., Cantalapiedra, C. P., García-Pereira, M. J., Gordon, S. P., Vogel, J. P., Igartua, E., et al. (2017). Analysis of plant pan-genomes and transcriptomes with GET_HOMOLOGUES-EST, a clustering solution for sequences of the same species. *Front. Plant Sci.* 8:184. doi: 10.3389/fpls.2017.00184

Cooper, E. A., Brenton, Z. W., Flinn, B. S., Jenkins, J., Shu, S., Flowers, D., et al. (2019). A new reference genome for *Sorghum bicolor* reveals high levels of sequence similarity between sweet and grain genotypes: implications for the genetics of sugar metabolism. *BMC Genomics* 20:420. doi: 10.1186/s12864-019-5734-x

Cuevas, H. E., and Prom, L. K. (2020). Evaluation of genetic diversity, agronomic traits, and anthracnose resistance in the NPGS Sudan Sorghum Core collection. *BMC Genomics* 21:88. doi: 10.1186/s12864-020-6489-0

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Dolatabadian, A., Bayer, P. E., Tirnaz, S., Hurgobin, B., Edwards, D., and Batley, J. (2020). Characterization of disease resistance genes in the Brassica napus pangenome reveals significant structural variation. *Plant Biotechnol. J.* 18, 969–982. doi: 10.1111/pbi.13262

Evans, J., McCormick, R. F., Morishige, D., Olson, S. N., Weers, B., Hilley, J., et al. (2013). Extensive variation in the density and distribution of DNA polymorphism in sorghum genomes. *PLoS ONE* 8:e79192. doi: 10.1371/journal.pone.0079192

Fracasso, A., Trindade, L. M., and Amaducci, S. (2016). Drought stress tolerance strategies revealed by RNA-Seq in two sorghum genotypes with contrasting WUE. *BMC Plant Biol.* 16:115. doi: 10.1186/s12870-016-0800-x

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., et al. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* 51, 1044–1051. doi: 10.1038/s41588-019-0410-2

Golicz, A. A., Batley, J., and Edwards, D. (2016b). Towards plant pangenomics. *Plant Biotechnol. J.* 14, 1099–1105. doi: 10.1111/pbi.12499

Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H., Martinez, P. A., et al. (2016a). The pangenome of an agronomically important crop plant Brassica oleracea. *Nat. Commun.* 7:13390. doi: 10.1038/ncomms13390

Gordon, S. P., Contreras-Moreira, B., Woods, D. P., Des Marais, D. L., Burgess, D., Shu, S., et al. (2017). Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nat. Commun.* 8:2184. doi: 10.1038/s41467-017-02292-8

Guo, H., Jiao, Y., Tan, X., Wang, X., Huang, X., Jin, H., et al. (2019). Gene duplication and genetic innovation in cereal genomes. *Genome Res.* 29, 261–269. doi: 10.1101/gr.237511.118

Hart, G. E., Schertz, K. F., Peng, Y., and Syed, N. H. (2001). Genetic mapping of *Sorghum bicolor* (L.) Moench QTLs that control variation in tillering and other morphological characters. *Theor. Appl. Genet.* 103, 1232–1242. doi: 10.1007/s001220100582

Hirsch, C. N., Foerster, J. M., Johnson, J. M., Sekhon, R. S., Muttoni, G., Vaillancourt, B., et al. (2014). Insights into the maize pan-genome and pan-transcriptome. *Plant Cell* 26, 121–135. doi: 10.1105/tpc.113.119982

Hubisz, M. J., Falush, D., Stephens, M., and Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour*. 9, 1322–1332. doi: 10.1111/j.1755-0998.2009.02591.x

Hurgobin, B., Golicz, A. A., Bayer, P. E., Chan, C. K. K., Tirnaz, S., Dolatabadian, A., et al. (2018). Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. *Plant Biotechnol. J.* 16, 1265–1274. doi: 10.1111/pbi.12867

Jensen, S. E., Charles, J. R., Muleta, K., Bradbury, P. J., Casstevens, T., Deshpande, S. P., et al. (2020). A sorghum practical haplotype graph facilitates genome-wide imputation and cost-effective genomic prediction. *Plant Genome* 13:e20009. doi: 10.1002/tpg2.20009

Johnson, S. M., Lim, F. L., Finkler, A., Fromm, H., Slabas, A. R., and Knight, M. R. (2014). Transcriptomic analysis of *Sorghum bicolor* responding to combined heat and drought stress. *BMC Genomics* 15:456. doi: 10.1186/1471-2164-15-456

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol*. 37, 907–915. doi: 10.1038/s41587-019-0201-4

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 14:R36. doi: 10.1186/gb-2013-14-4-r36

Kimani, W., Zhang, L.-M., Wu, X.-Y., Hao, H.-Q., and Jing, H.-C. (2020). Genome-wide association study reveals that different pathways contribute to grain quality variation in sorghum (Sorghum bicolor). *BMC Genomics* 21:112. doi: 10.1186/s12864-020-6538-8

Kong, L., Dong, J., and Hart, G. E. (2000). Characteristics, linkage-map positions, and allelic differentiation of *Sorghum bicolor* (L.) Moench DNA simple-sequence repeats (SSRs). *Theor. Appl. Genet*. 101, 438–448. doi: 10.1007/s001220051501

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Lasky, J. R., Upadhyaya, H. D., Ramu, P., Deshpande, S., Hash, C. T., Bonnette, J., et al. (2015). Genome-environment associations in sorghum landraces predict adaptive traits. *Sci. Adv*. 1:e1400218. doi: 10.1126/sciadv.1400218

Letunic, I., and Bork, P. (2019). Interactive Tree of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47, W256–W259. doi: 10.1093/nar/gkz239

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, R., Zhang, H., Zhou, X., Guan, Y., Yao, F., Song, G., et al. (2010). Genetic diversity in Chinese sorghum landraces revealed by chloroplast simple sequence repeats. *Genet. Resour. Crop Evol*. 57, 1–15. doi: 10.1007/s10722-009-9446-y

Lin, K., Zhang, N., Severing, E. I., Nijveen, H., Cheng, F., Visser, R. G. F., et al. (2014). Beyond genomic variation - comparison and functional annotation of three Brassica rapa genomes: a turnip, a rapid cycling and a Chinese cabbage. *BMC Genomics* 15:250. doi: 10.1186/1471-2164-15-250

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., et al. (2012). GAPIT: Genome association and prediction integrated tool. *Bioinformatics* 28, 2397–2399. doi: 10.1093/bioinformatics/bts444

McCormick, R. F., Truong, S. K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., et al. (2018). The *Sorghum bicolor* reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *Plant J.* 93, 338–354. doi: 10.1111/tpj.13781

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 28, 1297–1303. doi: 10.1101/gr.107524.110

Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H. T., Chan, C. K. K., et al. (2017). The pangenome of hexaploid bread wheat. *Plant J*. 90, 1007–1013. doi: 10.1111/tpj.13515

Morris, G. P., Ramu, P., Deshpande, S. P., Hash, C. T., Shah, T., Upadhyaya, H. D., et al. (2013). Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc. Natl. Acad. Sci. U.S.A*. 110, 453–458. doi: 10.1073/pnas.1215985110

Motlhaodi, T., Geleta, M., Bryngelsson, T., Fatih, M., Chite, S., and Ortiz, R. (2014). Genetic diversity in ex-situ conserved sorghum accessions of Botswana as estimated by: microsatellite markers. *Aust. J. Crop Sci*. 8, 35–43. Available online at: http://www.cropj.com/motlhaodi_8_2014_35_43.pdf

Obilana, A. B., Rao, E. P., Mangombi, N., and House, L. R. (1996). Classification of sorghum races in the Southern Africa Sorghum germplasm. *Sadc/Icrisat* 113–118. Available online at: http://oar.icrisat.org/4740/

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289–290. doi: 10.1093/bioinformatics/btg412

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., et al. (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457, 551–556. doi: 10.1038/nature07723

Peng, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2012). IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28, 1420–1428. doi: 10.1093/bioinformatics/bts174

Price, H. J., Dillon, S. L., Hodnett, G., Rooney, W. L., Ross, L., and Johnston, J. S. (2005). Genome evolution in the genus Sorghum (Poaceae). *Ann. Bot*. 95, 219–227. doi: 10.1093/aob/mci015

Ritter, K. B., McIntyre, C. L., Godwin, I. D., Jordan, D. R., and Chapman, S. C. (2007). An assessment of the genetic relationship between sweet and grain sorghums, within *Sorghum bicolor* ssp. *bicolor (L.)* Moench, using AFLP markers. *Euphytica* 157, 161–176. doi: 10.1007/s10681-007-9408-4

Sapkota, S., Boyles, R., Cooper, E., Brenton, Z., Myers, M., and Kresovich, S. (2020). Impact of sorghum racial structure and diversity on genomic prediction of grain yield components. *Crop Sci*. 60, 132–148. doi: 10.1002/csc2.20060

Saxena, R. K., Edwards, D., and Varshney, R. K. (2014). Structural variations in plant genomes. *Brief. Funct. Genomics* 13, 296–307. doi: 10.1093/bfgp/elu016

Schatz, M. C., Maron, L. G., Stein, J. C., Hernandez Wences, A., Gurtowski, J., Biggers, E., et al. (2014). Whole genome *de novo* assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. *Genome Biol*. 15:506. doi: 10.1186/PREACCEPT-27848725212 77375

Smit, A. F. A., Hubley, R., and Green, P. (2000). RepeatMasker. *Biotech Software Internet Rep*. 1, 36–39. doi: 10.1089/152791600319259

Tian, T., You, Q., Zhang, L., Yi, X., Yan, H., Xu, W., et al. (2016). SorghumFDB: sorghum functional genomics database with multidimensional network analysis. *Database* 2016:baw099. doi: 10.1093/database/baw099

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol*. 28, 511–515. doi: 10.1038/nbt.1621

Upadhyaya, H. D., Reddy, K. N., Vetriventhan, M., Gumma, M. K., Irshad Ahmed, M., Manyasa, E., et al. (2017). Geographical distribution, diversity and gap analysis of East African sorghum collection conserved at the ICRISAT genebank. *Aust. J. Crop Sci*. 11, 424–437. doi: 10.21475/ajcs.17.11.04.pne330

Usha Kiranmayee, K. N. S., Hash, C. T., Sivasubramani, S., Ramu, P., Amindala, B. P., Rathore, A., et al. (2020). Fine-mapping of sorghum stay-green qtl on chromosome10 revealed genes associated with delayed senescence. *Genes* 11, 2–26. doi: 10.3390/genes11091026

Valluru, R., Gazave, E. E., Fernandes, S. B., Ferguson, J. N., Lozano, R., Hirannaiah, P., et al. (2019). Deleterious mutation burden and its association with complex traits in sorghum (Sorghum bicolor). *Genetics* 211, 1075–1087. doi: 10.1534/genetics.118.301742

Varoquaux, N., Cole, B., Gao, C., Pierroz, G., Baker, C. R., Patel, D., et al. (2019). Transcriptomic analysis of field-droughted sorghum from seedling to maturity reveals biotic and metabolic responses. *Proc. Natl. Acad. Sci. U.S.A*. 116, 27124–27132. doi: 10.1073/pnas.1907500116

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., et al. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* 557, 43–49. doi: 10.1038/s41586-018-0063-9

Woldesemayat, A. A., and Ntwasa, M. (2018). Pathways and network based analysis of candidate genes to reveal cross-talk and specificity in the sorghum (*Sorghum bicolor* (L.) Moench) responses to drought and it's co-occurring stresses. *Front. Genet*. 9:557. doi: 10.3389/fgene.2018.00557

Yan, S., Wang, L., Zhao, L., Wang, H., and Wang, D. (2018). Evaluation of genetic variation among sorghum varieties from southwest China via genome resequencing. *Plant Genome* 11:170098. doi: 10.3835/plantgenome2017.11.0098

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P. C., Hu, L., et al. (2016). Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nat. Genet*. 48, 927–934. doi: 10.1038/ng.3596

Yao, W., Li, G., Zhao, H., Wang, G., Lian, X., and Xie, W. (2015). Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome Biol*. 16:187. doi: 10.1186/s13059-015-0757-3

Zhao, J., Bayer, P. E., Ruperao, P., Saxena, R. K., Khan, A. W., Golicz, A. A., et al. (2020). Trait associations in the pangenome of pigeon pea (Cajanus cajan). *Plant Biotechnol. J*. 18, 1946–1954. doi: 10.1111/pbi.13354

Zheng, L. Y., Guo, X., Sen, H.e, B., Sun, L. J., and Peng, Y., Dong, S.S., et al. (2011). Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). *Genome Biol*. 12:R114. doi: 10.1186/gb-2011-12-11-r114

Zhou, P., Silverstein, K. A. T., Ramaraj, T., Guhlin, J., Denny, R., Liu, J., et al. (2017). Exploring structural variation and gene family architecture with *de novo* assemblies of 15 Medicago genomes. *BMC Genomics* 18:261. doi: 10.1186/s12864-017-3654-1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.