A chromosome-scale genome sequence of sudangrass (Sorghum sudanense) highlights the genome evolution and regulation of dhurrin biosynthesis

Jieqin Li
Anhui Science and Technology University
Lihua Wang
Anhui Science and Technology University
Paul W. Bible
Marian University
Wenmiao Tu
Anhui Science and Technology University
Jian Zheng
Anhui Science and Technology University
Peng Jin
Anhui Science and Technology University
Yanlong Liu
Anhui Science and Technology University
Junli Du
Anhui Science and Technology University
Jiacheng Zheng
Anhui Science and Technology University
Yi-Hong Wang
University of Louisiana at Lafayette
Qiuwen Zhan (qwzhan@163.com)
Anhui Science and Technology University https://orcid.org/0000-0001-5491-7541

Research Article

Keywords: Sorghum sudanese, Genome, Genome evolution, GWAS, Dhurrin biosynthesis

Posted Date: July 6th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1777118/v1

Abstract

Sudangrass [Sorghum sudanense (Piper) Stapf] is a hybrid between grain sorghum and its wild relative S. bicolor ssp. verticilliflorum and is grown as a forage crop due to its high biomass production and low dhurrin content compared to sorghum. In this study, we sequenced the sudangrass genome and showed that the assembled genome was 715.95 Mb with 35,243 protein-coding genes. Phylogenetic analysis with whole genome proteomes demonstrated that the sudangrass genome was more similar to US commercial sorghums than to its wild relatives and cultivated sorghums from Africa. We confirmed that at seedling stage, sudangrass accessions contained significantly lower dhurrin as measured by hydrocyanic acid potential (HCN-p) than cultivated sorghum accessions. Genome-wide association study identified a QTL most tightly associated with HCN-p and the linked SNPs were located in the 3' UTR of Sobic.001G012300 which encodes CYP79A1, the enzyme that catalyzes the first step of dhurrin biosynthesis. As in other grasses such as maize and rice, we also found that copia/gypsy long terminal repeat retrotransposons were more abundant in cultivated than in wild sorghums, implying that crop domestication in the grasses was accompanied by increased copia/gypsy LTR retrotransposon insertions in the genomes.

Key Message

Sudangrass is more similar to US commercial sorghums than to cultivated sorghums from Africa sequence-wise and contain significantly lower dhurrin than sorghums. CYP79A1 is linked to dhurrin content in sorghum.

Introduction

Sudangrass [*Sorghum sudanense* (Piper) Stapf, also known as *S. bicolor* ssp. *drummondii*] is a hybrid between grain sorghum and its wild relative *S. bicolor* ssp. *Verticilliflorum* (Wiersema et al. 2007) and has been grown as a forage crop together with forage sorghums (Beck et al. 2013). However, when used as forage, sorghum poses a risk of cattle poisoning by dhurrin [p-hydroxy-(S)-mandelonitrile-β-D-glucoside], the precursor of hydrocyanic acid (HCN) (Hayes et al. 2015). Dhurrin is produced by sorghum and related species mainly in the leaves with much lower concentration in the stems and panicles and is measured by hydrocyanic acid potential (HCN-p) (De et al. 2011). Sudangrass contains lower concentration of dhurrin than sorghum-sudangrass hybrids, forage sorghum, shattercane, Johnsongrass, grain sorghum and sorghum almum (hybrid between grain sorghum and Johnsongrass) (De et al. 2011; Gorz et al. 1977; Loyd et al. 1970; McBee et al. 1980; Provin et al. 2012). Because of its low dhurrin content and high forage yield, sudangrass has been used to produce sudangrass-sorghum hybrids which display high forage biomass heterosis (Lu et al. 2011; Zhan et al. 2004). Within the cultivated sorghum, there is tremendous variation in dhurrin content in the leaves with a 17-fold difference between the highest and lowest dhurrin-containing entries (Hayes et al. 2015). Dhurrin plays a role of defense against herbivores and pathogens (Gleadow et al. 2014).

In sorghum, dhurrin is synthesized from L-tyrosine in multiple steps catalyzed by two multifunctional cytochrome P450 enzymes (CYP79A1 and CYP71E1) and a family 1 UDP-glucosyltransferase (UGT85B1), together with the P450 redox partner NADPH-dependent cytochrome P450 oxidoreductase (POR) (Gleadow et al. 2014; Laursen et al. 2016). The three genes coding for CYP79A1, CYP71E1 and UGT85B1 are clustered on chromosome 1 (Hayes et al. 2015; Gleadow et al. 2014; Darbani et al. 2016). In addition, the cluster also contains a membrane transporter (*SbMATE2*) (Darbani et al. 2016) and a gluthathione S-transferase (GST) which may participate in dhurrin turnover (Gleadow et al. 2014). On chromosome 8, there are two β-glucosidases that hydrolyze dhurrin, the dhurrinases *Dhr1* and *Dhr2* (Hayes et al. 2015; Cicek et al. 1998).

In this study, we sequenced and assembled the whole genome of sudangrass S722, characterized the evolution of the sudangrass genome. We also measured HCN-p in the mini core (Upadhyaya et al. 2009) and sudangrass accessions, performed genome-wide association studies (GWAS) and identified a strongly associated quantitative trait locus (QTL) which included two SNP and one indel marker. These markers were all located in the 3’ UTR of Sobic.001G012300 coding for CYP79A1, the enzyme that catalyzes the first step of dhurrin synthesis (Gleadow et al. 2014; Laursen et al. 2016).

Materials And Methods

Genome sequencing and assembly

DNA of sudangrass S722 was isolated using DNAsecure Plant Kit (Qiagen, Cat.No. DP320). After DNA isolation, 20 kb DNA fragments were recovered using the BluePippin system (Sage Science, Beverly, MA). A sequencing library was constructed with SQK-LSK109 kit and sequenced using PromethION 48 (Oxford Nanopore Technologies-ONT, UK). For short-read sequencing, DNA size of 350 bp were generated using the Covaris E210 Ultrasonicator (Woburn, MA) and the library was sequenced with BGI’s PCR DNBSEQ™ (Shenzhen, China). Similarly, a 350 bp Hi-C sequencing library was constructed and sequenced using an MGI-2000 Sequencer (Shenzhen, China). After sequencing, the short-read sequencing data were processed using SOAPnuke1.5.6 (Chen et al. 2018) with the following setting to obtain clean reads:"-n 0.01 -l 20 -q 0.1 -i -Q 2 -G -M 2 -A 0.5 -d".

For genome assembly, the long-read ONT sequencing data were *de novo* assembled into contigs using NECAT assembler (Chen et al. 2021). Three rounds of correction were applied subsequently using Racon v1.3.3 (Vaser et al. 2017) before Pilon (Walker et al. 2014) was used to improve the accuracy of the genome assembly by integrating the short-read sequencing data. Lastly, Hi-C technology (Lieberman-Aiden et al. 2009) was used to anchor primary contigs to pseudo-molecules and remove redundancy.

Genome annotation

For genome annotation, repeat sequences were identified using RepBase v21.12 (Bao et al. 2015), RepeatMasker v4.0.7, RepeatProteinMask v4.0.7, RepeatScout and RepeatModeler (http://www.repeatmasker.org), Piler (Edgar et al. 2005), Tandem Repeats Finder v4.09 (Benson 1999) or

LTR_FINDER v1.06 (Xu et al. 2007). Gene annotations based on RNA-seq data and de novo prediction. Genewise v2.4.1 (Birney et al. 2004), Augustus (Stanke et al. 2008), GlimmerHMM (Majoros et al. 2004), SNAP (http://homepage.mac.com/iankorf/), Genscan (Burge et al. 1997), FgeneSH (Salamov et al. 2000) and geneid v1.4.4 (Alioto et al. 2018) were used to predicate and annotate genes. tRNAscanSE software (Chan et al. 2019) was used to identify tRNAs using default settings. INFERNAL software was used to determine miRNA and snRNA against Rfam database (https://rfam.xfam.org/).

Gene family analysis

The proteomes of sudangrass and three other sorghum genus species were used to cluster gene families. OrthoFinder (Emms et al. 2015) was used to identify protein paralogs and orthologs sequences among the four species with the default parameters. The Venn graph was drawn using OrthoVenn2 (Xu et al. 2019).

Phylogenetic analysis

Single-gene families shared by sudangrass and eight other plant species were identified by OrthoFinder. RAxML was used to construct the maximum likelihood phylogenetic tree using the GTRGAMMA model with *Arabidposis* thaliana as an out group with a bootstrap value of 1000. The MCMCtree program of PAML (Yang 2007) was used to estimate the divergence time. The times of divergence between *Arabidposis thaliana* and *Oryza sativa* (115–308 Mya), and *Oryza sativa* and *Zea mays* (42–52 Mya) were from the TimeTree database (http://timetree.org/).

To explore the evolutionary history of the gene families, gene expansion and contraction analysis was performed using CAFE 5 (Mendes et al. 2021) with default parameters. Gene trees calculated by Orthofinder were retrieved, and gene families were extracted based on sequence similarity using Orthofinder.

Proteomes were collected for 17 sequenced plant varieties. These included sudangrass (S722), the *Sorghrum bicolor* reference (Sbicolor proteome v3.1.1) (Goodstein et al. 2012), S. *bicolor* Rio (v2.1, Phytozome) (Cooper et al. 2019), S. *bicolor* RTx430 (v2.1, Phytozome) (Deschamps et al. 2018), S. *bicolor* BTx642 (v1.1, Phytozome), SC187 (v1.1, Phytozome), *S. bicolor ssp. bicolor* durra (IS929) (Tao et al. 2021), S. *bicolor ssp. bicolor* Guinea /conspicuum (IS3614) (Tao et al. 2021), S. *bicolor* ssp. *bicolor* Kafir (IS8525) (Tao et al. 2021), S. *bicolor* ssp. *bicolor* Caudatum /zerazera (IS12661) (Mitros et al. 2020), S. *bicolor* ssp. *verticilliflorum* (PI536008) (Tao et al. 2021), S. *bicolor* ssp. *drummondii* (PI532566) (Tao et al. 2021), S. *bicolor* ssp. *bicolor* Margaritiferum (IS19953) (Tao et al. 2021), another variety of *S. bicolor* ssp. *verticilliflorum* (AusTRCF317961) (Tao et al. 2021), the wild relative *S. propinquum* (S369-1) (Tao et al. 2021), *Miscanthus sinensis* (v7.1, Phytozome) (Mitros et al. 2020), and the outgroup *Zea mays* reference (v4, Phytozome). For each variety, the longest isoform amino acid sequences were selected and processed using the Orthofinder pipeline to detect orthogroups and orthologs (Emms et al. 2015; Emms et al. 2019). For orthogroup sequence alignments, the '-M msa' option was used (orthofinder -t 8 -a 8 -M msa -f prots_folder/). The '-M msa' option invokes a more sensitive sequence alignment using MAFFT (Katoh et al. 2013) v 7.481. Gene and species trees were constructed by Orthofinder using FastTree (Price et al.

2010) version 2.1.11 from multiple sequence alignments. The phylogenetic relationship between varieties (called the “species tree”), was inferred by Orthofinder based on conserved single copy proteins using the STRIDE (Emms et al. 2017) and STAG (Emms et al. 2018) algorithms.

Chromosomal structural change tracking

To assess global structural changes, sorghum genome was compared to the sudangrass genome using CHROMEISTER (Pérez-Wohlfeil et al. 2019) (command: “CHROMEISTER -query genome1.fa -db genome2.fa -out outputfile”). The generated dot-plot was drawn using the compute_score.R R script provided by CHROMEISTER to create image files. Before analysis, each genome assembly was filtered to use only its chromosomes. Any genome with only scaffold-level assembly quality was filtered to its 10 largest scaffolds (all > 19Mb). Large scale structural changes and their corresponding coordinates were calculated using CHROMEISTER‘s detect_events.py script (command: “detect_events.py outputfile.raw.txt png”).

LTR analysis

LTR_FINDER and LTRharvest (Xu et al. 2007) were used to search intact long terminal repeat retrotransposons (LTR-RTs) against the genome sequences. LTR-RT insertion time was calculated using T = K/2r (Steinbiss et al. 2009). All LTR-RTs were grouped into Ty1/*copia*, Ty3/*gypsy* or other superfamilies based on their structure and proteins domains using LTRdigest (Ossowski et al. 2010). The density graph of LTR-RTs for the Sorghum genus was drawn using ggplot2 in R.

HCN-p and GWAS

Sorghum mini core and seven sudangrass accessions were planted in plastic trays in triplicate. The plants were grown at 28 ℃ under 12 hours light and 12 hours dark photoperiod. Two weeks after planting, the first leaf was sampled to measure HCN-p as described by Gorz et al. (1977) (Gorz et al. 1977).

GWAS was performed according to Wang et al. (2021) (Wang et al. 2021) using 6,094,317 SNPs and 957,449 indels. Kinship matrix (K) was generated using EMMAX (Kang et al. 2010). GWAS analyses were performed using EMMAX with Q matrix. The modified Bonferroni correction was used to determine the genome-wide significance thresholds of the GWAS, based on a nominal level of $\alpha = 0.05$ corresponding raw P values of 8.2×10^{-9} or $-\log_{10}(P)$ values of 8.08. Candidate genes were identified using the reference sequence *Sorghum bicolor* v3.1.1 curated at Phytozome 13 (https://phytozome-next.jgi.doe.gov/).

Results

Genome sequencing and assembly

On the basis of K-mer distribution assessment (K = 25), the estimated genome size of sudangrass (2n = 20) was 741.34 Mb with heterozygosity of 0.225% and repeat contents of 58.4% (Table 1, Supplementary

Table 1, Supplementary Fig. 1). DNBseq, ONT, and Hi-C data were combined to yield a high-quality and chromosome-level reference genome. In total, 106.32 Gb of ONT long reads (~ 143.41x coverage of the genome), 113.78 Gb (~ 153.55x coverage of the genome) of DNB clean reads, 107.5Gb (~ 145.07x coverage of the genome) of Hi-C data were produced, resulting in 442–fold coverage of the genome. The assembled genome was 715.95 Mb with scaffold N50 of 71.60 Mb and contig N50 of 28.97 Mb (Table 1). The longest contig and scaffold were 44.15 Mb and 86.04 Mb, respectively. The Hi-C data were used to order and anchor the assembled sequences onto the 10 chromosomes (Supplementary Fig. 2). The genome contained 44.33% GC.

The mapping rate was 95.72% when BGI sequencing reads and the assembled genome were aligned. BUSCO analysis was carried out to evaluate the quality of the assembled genome and showed that sudangrass assembly contained 97.9% of the highly conserved genes common across eukaryotes (Supplementary Table 2). In addition, 142 syntenic blocks and 1,489 paralog groups were identified based on self-alignment of 35,243 annotated genes, indicating that the sudangrass genome has undergone frequent segmental duplications and interchromosome fusions in its evolutionary history (Fig. 1).

Table 1
Assembly and annotation of the sudangrass genome

Estimated genome size (Mb)	**741.34**
Assembled genome size (Mb)	720.10
Number of Scaffolds	10
Number of N50 Scaffolds	5
Number of contigs	115
Number of N50 contigs	11
Longest Chr (Mb)	86.04
GC content (%)	44.33
Transposable elements (%)	67.33
Predicated protein-coding genes	35,243
Average gene length (bp)	3316.36
Average exon length (bp)	282.50

Genome annotation

RNA sequence and homology searches were used to identify protein-coding genes in the sudangrass genome. Overall, 35,243 protein-coding genes were identified. On average, gene length was 3.32 kb and exon length was 282.50 bp (Supplementary Table 3). A total of 90.25% (31,086 genes) of the 35,243genes were annotated by homology to known proteins, domains or transcripts (Supplementary

Table 4). In total, 1647 transcription factors (TFs) in 56 TF families were identified in the genome. These included 188 MYB, 160 bHLH and 94 WRKY (Supplementary Table 5).

The sudangrass genome contained 510.09 Mb repetitive sequences (71.25% of the genome) using homology-based and de novo methods, of which 3.68% were tandem repeat and interspersed repeats (26.36 Mb). Long terminal repeats (LTRs) retrotransposons was the most abundant interspersed repeats, representing 57.1% of the genome (394.71 Mb), followed by DNA transposons at 10.12% (72.48 Mb). The non-LTR retrotransposons LINEs (long interspersed nuclear elements) and SINEs (short interspersed nuclear elements) accounted for 2.08% (14.88 Mb) of the genome (Supplementary Table 6). Genes annotated as ncRNAs included 2,751 miRNAs, 690 tRNAs, 672 rRNAs, and 4,670 snRNAs (Supplementary Table 7).

Comparison of gene families

A gene family cluster evaluation from the whold genomes of four *Sorghum* genus (*S. bicolor, S. proniquum, S. bicolor* ssp. *verticilliflorum* and sudangrass) was performed. The four genomes shared 17,155 gene families and 448 gene families were sudangrass-specific (Fig. 2A). These 448 gene families consisting of 3,141 genes were used to perform GO analysis (cutoff < 0.05). These genes were enriched in cell component (endoplasmic reticulum lumen) and 15 molecular function categories (ATPase-coupled cation transmembrane transporter activity, metal ion binding etc.) (Supplementary Fig. 3 and Supplementary Table 8).

The CAFE program was used to perform gene expansion and contraction analysis. Comparative genomic analyses were performed among nine representative plant species (Fig. 2B) and 276 and 905 gene family expansion and contractions, respectively, were found in sudangrass, after its divergence from S. bicolor, indicating that more sudangrass gene families have undergone contraction than expansion. Among the gene families, 638 significantly expanded genes and 603 significantly contracted genes were found at the 0.05 level. KEEG analysis showed that expanded genes were enriched in 45 pathways including acridone alkaloid biosynthesis, flavonoid biosynthesis and circadian rhythm (Supplementary Table 9 and Supplementary Fig. 4). Contracted genes were enriched in 17 pathways including isoflavonoid biosynthesis and stilbenoid, diarylheptanoid and gingerol biosynthesis etc. (Supplementary Table 10 and Supplementary Fig. 5).

A divergence tree was constructed based on expansion-contraction of 186 single-copy orthologs. The results indicated that the *Sorghum* genus diverged from Z. *mays* about 21.6 Mya (16.2–27.4), and sudangrass diverged from *S. bicolor* about 1.1 Mya (0.6–1.7) (Fig. 2B). We also constructed a phylogenetic tree using 17 sequenced plant varieties. In this phylogenetic tree, sudgangrass (S722) was most closely related to the five cultivated sorghums BTx623, Rio, SC187, RTx430 and BTx642 and distinct from wild relatives and cultivated sorghums from Africa (Supplementary Fig. 6).

Chromosome changes compared to the sorghum genome

Comparisons of the sudangrass genome with the sorghum reference genome by chromosome using Chromeister (Pérez-Wohlfeil et al. 2019) indicated largely collinearity between the two, except between sudangrass chromosome 4 and the reference chromosome 5 (Fig. 3). In this case, while the middle of the two chromosomes was largely colinear, both arms were inverted (Fig. 3). Between sudangrass chromosome 1 and the reference chromosome 1 and between sudangrass chromosome 6 and the reference chromosome 7, it was the middle portion of the chromosomes that were inverted (Fig. 3).

The LTR analysis

Transposons, especially LTRs, are important to the evolution of genome structure. LTRs were identified and compared among the four sorghum species. In total, 3076, 7019, 6005 and 1940 intact LTRs were found in sudangrass, *S. bicolor, S. bicolor* ssp. *verticilliflorum* and *S. proniquum*, respectively (Fig. 4A). There was no significant proliferation of LTRs in the last 2 Mya for *S. proniquum*. But for *S. bicolor, S.bicolor* ssp. *verticilliflorum* and sudangrass, there has been continuous and substantial LTR accumulation in the last 2 Mya (Supplementary Fig. 6). For *S. bicolor* and *S. bicolor* ssp. *verticilliflorum*, LTR number expanded three times more than *S. proniquum*. For sudangrass, LTR number only expanded 0.5 times more than *S. proniquum*.

Approximately 81.21%, 81.72%, 79.30% and 84.03% of the intact LTRs had at least one protein domain in the four species (*S. bicolor, S. bicolor* ssp. *verticilliflorum, S. proniquum*, and sudangrass), respectively. There was no significant difference among the four species in the number of LTR with protein domain. As in other species, the majority of LTRs were Ty1/*copia* or Ty3/*gypsy*. Specifically, 70.96%, 69.69%, 64.14% and 59.92% were Ty3/*gypsy* in *S. bicolor*, *S. bicolor* ssp. *verticilliflorum*, *S. proniquum* and sudangrass, respectively, and the respective number for Ty1/*copia* were 10.76%, 9.61%, 19.89% and 21.29%. Compared to *S. proniquum*, the number of Ty1/*copia* was 0.5, 0.7 and 1.0 times more in *S. bicolor* ssp. *verticilliflorum*, *S. bicolor* and sudangrass, respectively. But for Ty3/*gypsy*, it was 2.4, 3.0 and 1.48 times more in the three species, respectively, compared to that in *S. proniquum*.

We also compared the expansion curve in the four species (Fig. 4B and 4C). For Ty3/*gypsy*, there was no sharp peak in *S. proniquum*. For sudangrass, *S. bicolor* and *S. bicolor* ssp. *verticilliflorum*, there were a burst in 1.06, 0.22 and 0.35 Mya for Ty3/*gypsy* and a burst in 0.32, 0.12 and 0.21 Mya for Ty1/*copia*, respectively. These showed that *S. bicolor* and *S. bicolor* ssp. *verticilliflorum* had similar expansion pattern and sudangrass had distinct expansion pattern compared to the other two species (Fig. 4B and 4C).

HCN-p in mini core/sudangrass accessions and GWAS

We measured leaf HCN-p (in ppm) in 2-week-old seedlings of 227 mini core sorghum and seven sudangrass accessions including S722. HCN-p ranged from 189–717 with an average of 381 in the mini core accessions while in the sudangrass accessions it ranged from 71–233 with an average of 140. Overall, sorghum accessions had higher HCN-p than sudangrass accessions ($p = 0.0000383$; Fig. 5).

We performed GWAS with the mini core HCN-p data using 6,094,317 SNPs and 957,449 indels. No markers were associated with HCN-p with $-\log_{10}(P)$ value greater than the threshold of 8.08 but we found one QTL with the highest $-\log_{10}(P)$ value (7.94 and 7.47 for the two SNPs and 7.73 for the indel) mapped to chromosome 1 (Fig. 6A and 6B), in the same gene cluster mapped by Hayes et al. (2015). The two SNPs (Fig. 6C) and one indel (Fig. 6D) with the strongest association with HCN-p were all located in the 3' UTR of Sobic.001G012300 which encodes CYP79A1 (Fig. 6E), the enzyme that catalyzes the first committed step of dhurrin biosynthesis, converting L-tyrosine into (Z)-p-hydroxyphenylacetaldehyde oxime (Gleadow et al. 2014; Laursen et al. 2016).

Discussion

In this study, we found that sudgangrass (S722) is most closely related to the five sequenced US commercial sorghums BTx623, Rio, SC187, RTx430 and BTx642 (Supplementary Fig. 6) which all produced panicles under field conditions. Although BTx623, SC187, RTx430 and BTx642 are all known grain sorghum varieties, Rio as a sweet sorghum variety can also produce grain yield comparable to BTx623 in some environments in addition to its high biomass and sugar content (Murray et al. 2008). In contrast, sudangrass produces lower grain yield (Li JQ unpublished results) and in this aspect sudangrass is not as fully domesticated.

We analyzed 227 sorghum mini core and seven sudangrass accessions for HCN-p and found that sorghum accessions contained higher HCN-p than sudangrass accessions. This is in agreement with previous studies that showed lower leaf HCN-p in sudangrass compared to sorghum (Gorz et al. 1977; Loyd et al. 1970; McBee et al. 1980). For example, Gorz et al. (1977) analyzed nine grain sorghum A/R lines, six sweet sorghums, 13 forage sorghums, and nine sudangrass varieties with average/(range) HCN-p of 1027/(861 ~ 1208), 1130/(976 ~ 1341), 983/(80 ~ 1560), and 445/(286 ~ 614) ppm, respectively.

There seems to be a correlation between the degree of domestication and the copy number of *copia/gypsy* LTR retrotransposons. In the sudangrass and the five US commercial sorghum clade shown in Supplementary Fig. 6, the copy number was 3076 for sudangrass S722 and 7019, 5844, 7203, 7308, and 7467 for BTx623, Rio, RTx430, SC187, and BTx642, respectively. Among the five varieties, the four commercial grain sorghums (BTx623, RTx430, SC187, and BTx642) contain similar copy number, but Rio has lower copy number (Supplementary Fig. 6) and is grown as sweet sorghum for its high stem juice sugar content (Murray et al. 2008). All other cultivated sorghums or wild relatives have copy number lower than the four grain sorghums (Supplementary Fig. 6). Our results confirm findings by other studies in other grasses. For example, in cultivated *indica* and *japonica* rice *gypsy* accounted for 21% and 22% of their respective genomes while this number ranges from 9–14% in wild rice relatives (Li et al. 2017). In another rice study, the 50 most abundant LTR retrotransposons numbered from 1885 ~ 3224 in cultivated *indica* and *japonica* rice and 337 ~ 1491 in wild rice relatives (Stein et al. 2018). Similarly in maize, in a study of 91 improved, 24 landrace and 10 teosinte maize accessions, Zhang and Qi (2019) (Zhang et al. 2019) found that landraces contained more *copia/gypsy* than teosinte (793,808 vs. 460,394) although improved maize contained far less than teosinte (154,136 vs. 460,394) on average. These results suggest

that crop domestication at least in the grasses was accompanied by increased *copia/gypsy* LTR retrotransposon insertions in the genomes.

The two SNPs (Fig. 6C) and one indel (Fig. 6D) with the strongest association with HCN-p were all located in the 3' UTR of Sobic.001G012300 which encodes CYP79A1 (Fig. 6E), the enzyme that catalyzes the first committed step of dhurrin biosynthesis, converting L-tyrosine into (Z)-p-hydroxyphenylacetaldehyde oxime (Gleadow et al. 2014; Laursen et al. 2016). The *CYP79A1* gene is critical to dhurrin biosynthesis as antisense plants reduces HCN from 221.4 μg/g to 76.2 μg/g (Pandey et al. 2019) and missense mutations P414L (Blomstedt et al. 2012) or C493Y (Skelton 2014) in the gene shuts down dhurrin production. One possible reason is that transcript levels of *CYP79A1* and *CYP71E1* are almost perfectly correlated (R = 0.956) (Choi et al. 2020) and CYP71E1 catalyzes the second of the three steps in dhurrin biosynthesis (Gleadow et al. 2014; Laursen et al. 2016). These GWAS results are based on HCN-p measurements of 2-week-old seedling leaves. Future studies may need to use data from seedling stage as well as other stages relevant to forage sorghum/sudangrass production.

In conclusion, we have sequenced a sudangrass genome producing a chromosome level assembly that is 715.95 Mb in size containing 35,243 genes. Phylogenetic analysis with whole genome proteomes also showed that the sudangrass genome was more similar to but distinct from the cultivated US sorghums. We confirmed that at seedling stage, sudangrass accessions contained significantly lower HCN-p than cultivated sorghum accessions. GWAS identified a QTL most tightly associated with HCN-p and the linked SNPs were located in the 3' UTR of Sobic.001G012300 which encodes CYP79A1, the enzyme that catalyzes the first step of dhurrin biosynthesis. These insights will guide future studies to interrogate the regulatory mechanisms of the dhurrin pathway.

Declarations

Acknowledgments

This work was supported by Anhui Provincial Natural Science Fund (2008085MC73), the National Natural Science Foundation of China (31971993), Anhui Provincial Key R&D Programme (202004b11020003) and the Key Project of Natural Science Research of Anhui Provincial Education Department (KJ2021ZD0108).

Conflict of interest

We declare that there are no conflicts of interest.

Author contribution

LW prepared samples for genome sequencing; WT, JZ, PJ, JD and JZ measured HCN-p in the seedlings of the mini core accessions and sudangrass; PWB performed chromosome alignment and phylogenetic analysis; YW performed statistic test for HCN-p between sudangrass and sorghums, and drafted the

manuscript; QZ designed the HCN-p experiment, JL and LW performed GWAS, phylogenetic and genome and transposon analyses. All authors reviewed and approved the final manuscript.

Data availability

The raw sequencing data of sudangrass have been deposited in NCBI with the Bioproject ID PRJNA831289. The assembled genome of sudangrass have been deposited in NCBI with Bioproject ID PRJNA830304.

References

1. Alioto T, Blanco E, Parra G, Guigó R (2018) Using geneid to Identify Genes. Curr Protoc Bioinformatics 64:e56. https://doi.org/10.1002/cpbi.56
2. Bao W, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 6:11. https://doi.org/10.1186/s13100-015-0041-9
3. Beck P, Poe K, Stewart B et al (2013) Effect of brown midrib gene and maturity at harvest on forage yield and nutritive quality of sudangrass. Grassl Sci 59:52–58. https://doi.org/10.1111/grs.12007
4. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580. https://doi.org/10.1093/nar/27.2.573
5. Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. Genome Res 14:988–995. https://doi.org/10.1101/gr.1865504
6. Blomstedt CK, Gleadow RM, O'Donnell N et al (2012) A combined biochemical screen and TILLING approach identifies mutations in Sorghum bicolor L. Moench resulting in acyanogenic forage production. Plant Biotechnol J 10:54–66. https://doi.org/10.1111/j.1467-7652.2011.00646.x
7. Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94. https://doi.org/10.1006/jmbi.1997.0951
8. Chan PP, Lowe TM (2019) tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. Methods Mol Biol 1962:1–14. https://doi.org/10.1007/978-1-4939-9173-0_1
9. Chen Y, Chen Y, Shi C et al (2018) SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. Gigascience 7:1–6. https://doi.org/10.1093/gigascience/gix120
10. Chen Y, Nie F, Xie S-Q et al (2021) Efficient assembly of nanopore reads via highly accurate and intact error correction. Nat Commun 12:60. https://doi.org/10.1038/s41467-020-20236-7
11. Choi SC, Chung YS, Lee YG et al (2020) Prediction of Dhurrin Metabolism by Transcriptome and Metabolome Analyses in Sorghum. Plants (Basel) 9:E1390. https://doi.org/10.3390/plants9101390
12. Cicek M, Esen A (1998) Structure and expression of a dhurrinase (beta-glucosidase) from sorghum. Plant Physiol 116:1469–1478. https://doi.org/10.1104/pp.116.4.1469
13. Cooper EA, Brenton ZW, Flinn BS 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. https://doi.org/10.1186/s12864-019-5734-x

14. Darbani B, Motawia MS, Olsen CE et al (2016) The biosynthetic gene cluster for the cyanogenic glucoside dhurrin in Sorghum bicolor contains its co-expressed vacuolar MATE transporter. Sci Rep 6:37079. https://doi.org/10.1038/srep37079
15. De Nicola GR, Leoni O, Malaguti L et al (2011) A simple analytical method for dhurrin content evaluation in cyanogenic plants for their utilization in fodder and biofumigation. J Agric Food Chem 59:8065–8069. https://doi.org/10.1021/jf200754f
16. Deschamps S, Zhang Y, Llaca V et al (2018) A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. Nat Commun 9:4844. https://doi.org/10.1038/s41467-018-07271-1
17. Edgar RC, Myers EW (2005) PILER: identification and classification of genomic repeats. Bioinformatics 21:i152–i158. https://doi.org/10.1093/bioinformatics/bti1003
18. Emms DM, Kelly S (2017) STRIDE: Species Tree Root Inference from Gene Duplication Events. Mol Biol Evol 34:3267–3278. https://doi.org/10.1093/molbev/msx259
19. Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157. https://doi.org/10.1186/s13059-015-0721-2
20. Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20:238. https://doi.org/10.1186/s13059-019-1832-y
21. Emms DM, Kelly S (2018) STAG: Species Tree Inference from All Genes. Evolutionary Biology
22. Gleadow RM, Møller BL (2014) Cyanogenic glycosides: synthesis, physiology, and phenotypic plasticity. Annu Rev Plant Biol 65:155–185. https://doi.org/10.1146/annurev-arplant-050213-040027
23. Goodstein DM, Shu S, Howson R et al (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40:D1178–1186. https://doi.org/10.1093/nar/gkr944
24. Gorz HJ, Haag WL, Specht JE, Haskins FA (1977) Assay of p-Hydroxybenzaldehyde as a measure of hydrocyanic acid potential. Agronomy & Horticulture – Faculty Publications.261. https://digitalcommons.unl.edu/agronomyfacpub/261
25. Hayes CM, Burow GB, Brown PJ et al (2015) Natural Variation in Synthesis and Catabolism Genes Influences Dhurrin Content in Sorghum. Plant Genome 8. https://doi.org/10.3835/plantgenome2014.09.0048. eplantgenome2014.09.0048
26. Kang HM, Sul JH, Service SK et al (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42:348–354. https://doi.org/10.1038/ng.548
27. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010
28. Laursen T, Borch J, Knudsen C et al (2016) Characterization of a dynamic metabolon producing the defense compound dhurrin in sorghum. Science 354:890–893.

https://doi.org/10.1126/science.aag2347

29. Li X, Guo K, Zhu X et al (2017) Domestication of rice has reduced the occurrence of transposable elements within gene coding regions. BMC Genomics 18:55. https://doi.org/10.1186/s12864-016-3454-z
30. Lieberman-Aiden E, van Berkum NL, Williams L et al (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326:289–293. https://doi.org/10.1126/science.1181369
31. Loyd RC, Gray E (1970) Amount and Distribution of Hydrocyanic Acid Potential during the Life Cycle of Plants of Three Sorghum Cultivars [1]. Agron J 62:394–397. https://doi.org/10.2134/agronj1970.00021962006200030025x
32. Majoros WH, Pertea M, Salzberg SL (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20:2878–2879. https://doi.org/10.1093/bioinformatics/bth315
33. McBee GG, Miller FR (1980) Hydrocyanic Acid Potential in Several Sorghum Breeding Lines as Affected by Nitrogen Fertilization and Variable Harvests [1]. Crop Sci 20:232–234. https://doi.org/10.2135/cropsci1980.0011183X002000020020x
34. Mendes FK, Vanderpool D, Fulton B, Hahn MW (2020) CAFE 5 models variation in evolutionary rates among gene families. https://doi.org/10.1093/bioinformatics/btaa1022. Bioinformatics btaa1022
35. Mitros T, Session AM, James BT et al (2020) Genome biology of the paleotetraploid perennial biomass crop Miscanthus. Nat Commun 11:5442. https://doi.org/10.1038/s41467-020-18923-6
36. Murray SC, Sharma A, Rooney WL et al (2008) Genetic Improvement of Sorghum as a Biofuel Feedstock: I. QTL for Stem Sugar and Grain Nonstructural Carbohydrates. Crop Sci 48:2165–2179. https://doi.org/10.2135/cropsci2008.01.0016
37. Ossowski S, Schneeberger K, Lucas-Lledó JI et al (2010) The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science 327:92–94. https://doi.org/10.1126/science.1180677
38. Pandey AK, Madhu P, Bhat BV (2019) Down-Regulation of CYP79A1 Gene Through Antisense Approach Reduced the Cyanogenic Glycoside Dhurrin in [Sorghum bicolor (L.) Moench] to Improve Fodder Quality. Front Nutr 6:122. https://doi.org/10.3389/fnut.2019.00122
39. Pérez-Wohlfeil E, Diaz-Del-Pino S, Trelles O (2019) Ultra-fast genome comparison for large-scale genomic experiments. Sci Rep 9:10274. https://doi.org/10.1038/s41598-019-46773-w
40. Price MN, Dehal PS, Arkin AP (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS ONE 5:e9490. https://doi.org/10.1371/journal.pone.0009490
41. Provin TL, Pitt JL, Texas A (2003)& M University System
42. Salamov AA, Solovyev VV (2000) Ab initio gene finding in Drosophila genomic DNA. Genome Res 10:516–522. https://doi.org/10.1101/gr.10.4.516

43. Skelton JL (2014) EMS induced mutations in dhurrin metabolism and their impacts on sorghum growth and development
44. Stanke M, Diekhans M, Baertsch R, Haussler D (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24:637–644. https://doi.org/10.1093/bioinformatics/btn013
45. Stein JC, Yu Y, Copetti D et al (2018) Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. Nat Genet 50:285–296. https://doi.org/10.1038/s41588-018-0040-0
46. Steinbiss S, Willhoeft U, Gremme G, Kurtz S (2009) Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Res 37:7002–7013. https://doi.org/10.1093/nar/gkp759
47. Tao Y, Luo H, Xu J et al (2021) Extensive variation within the pan-genome of cultivated and wild sorghum. Nat Plants 7:766–773. https://doi.org/10.1038/s41477-021-00925-x
48. Upadhyaya HD, Pundir RPS, Dwivedi SL et al (2009) Developing a Mini Core Collection of Sorghum for Diversified Utilization of Germplasm. Crop Sci 49:1769–1780. https://doi.org/10.2135/cropsci2009.01.0014
49. Vaser R, Sović I, Nagarajan N, Šikić M (2017) Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res 27:737–746. https://doi.org/10.1101/gr.214270.116
50. Walker BJ, Abeel T, Shea T et al (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE 9:e112963. https://doi.org/10.1371/journal.pone.0112963
51. Wang L, Upadhyaya HD, Zheng J et al (2021) Genome-Wide Association Mapping Identifies Novel Panicle Morphology Loci and Candidate Genes in Sorghum. Front Plant Sci 12:743838. https://doi.org/10.3389/fpls.2021.743838
52. Wen ZQ, Qiang QZ (2004) Heterosis utilization of hybrid between sorghum [Sorghum bicolor (L.) Moench] and sudangrass [Sorghum sudanense (Piper) Stapf]. Acta Agronomica Sinica
53. Wiersema JH, Dahlberg J (2007) The nomenclature of *Sorghum bicolor* (L.) Moench (*Gramineae*). Taxon 56:941–946. https://doi.org/10.2307/25065876
54. Xiao-ping L, Jin-feng Y, Cui-ping G, Acharya S (2011) Quantitative trait loci analysis of economically important traits in *Sorghum bicolor* × *S. sudanense* hybrid. Can J Plant Sci 91:81–90. https://doi.org/10.4141/cjps09112
55. Xu L, Dong Z, Fang L et al (2019) OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Res 47:W52–W58. https://doi.org/10.1093/nar/gkz333
56. Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35:W265–268. https://doi.org/10.1093/nar/gkm286
57. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591. https://doi.org/10.1093/molbev/msm088

58. Zhang X, Qi Y (2019) The Landscape of Copia and Gypsy Retrotransposon During Maize Domestication and Improvement. Front Plant Sci 10:1533. https://doi.org/10.3389/fpls.2019.01533

Figures

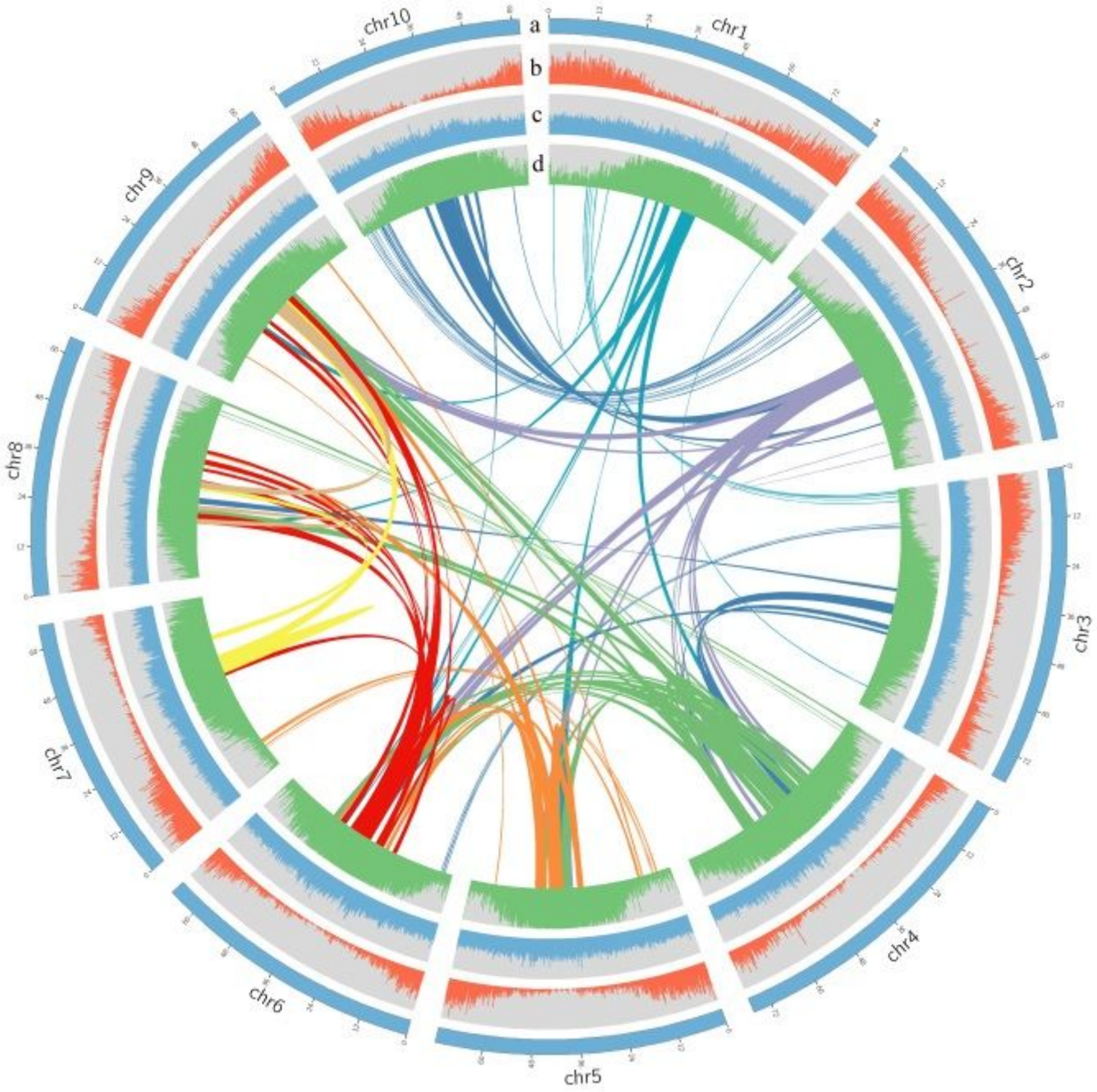

Figure 1

Genome features of sudangrass (*S. sudanense*). Track a, the ten chromosomes (in Mb scale). Track b, gene density. Track c, GC content. Track d, repeat density. Colored lines in the center represent interchromosomal synteny.

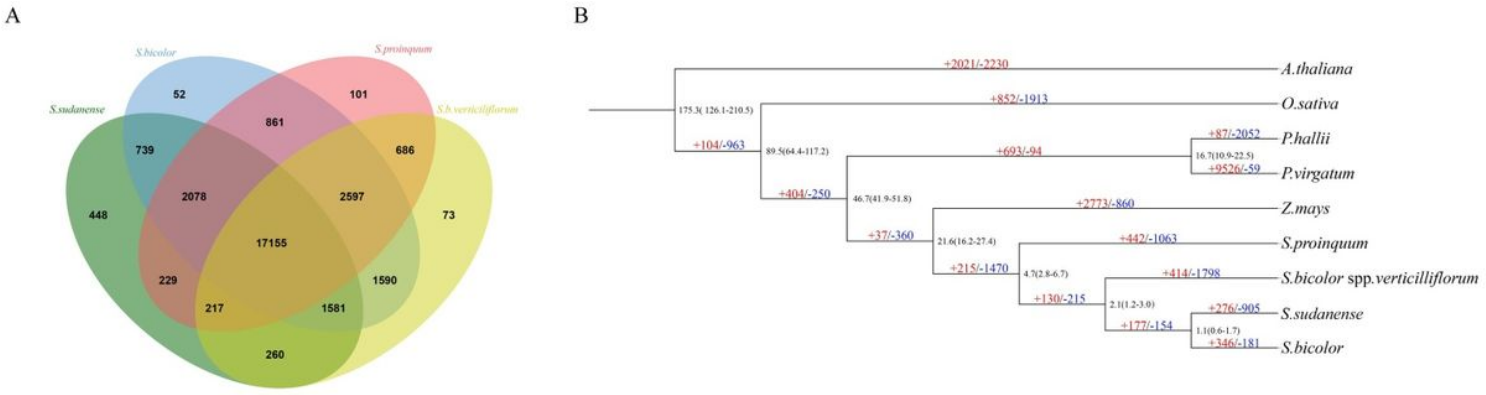

Figure 2

Comparison of gene families. A. Venn diagrams displaying the number of gene families shared among four sorghum species. B. Phylogenetic tree based on expansion-contraction of 186 single-copy orthologous genes. The divergence time is given in millions of years in node. The red and blue numbers on each branch presents expanded and contracted gene families, respectively.

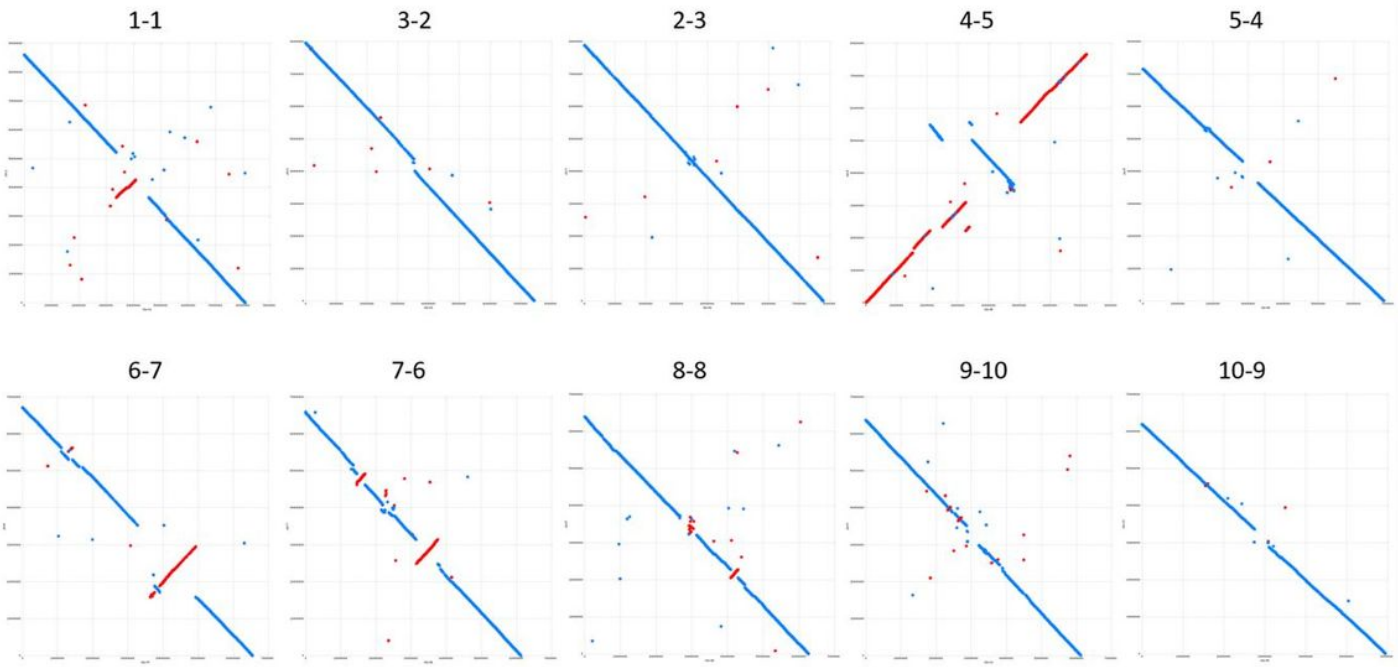

Figure 3

Chromosome structural comparison between sudangrass (first number) and the reference genome (second number) by Chromeister. Red lines indicate inversion.

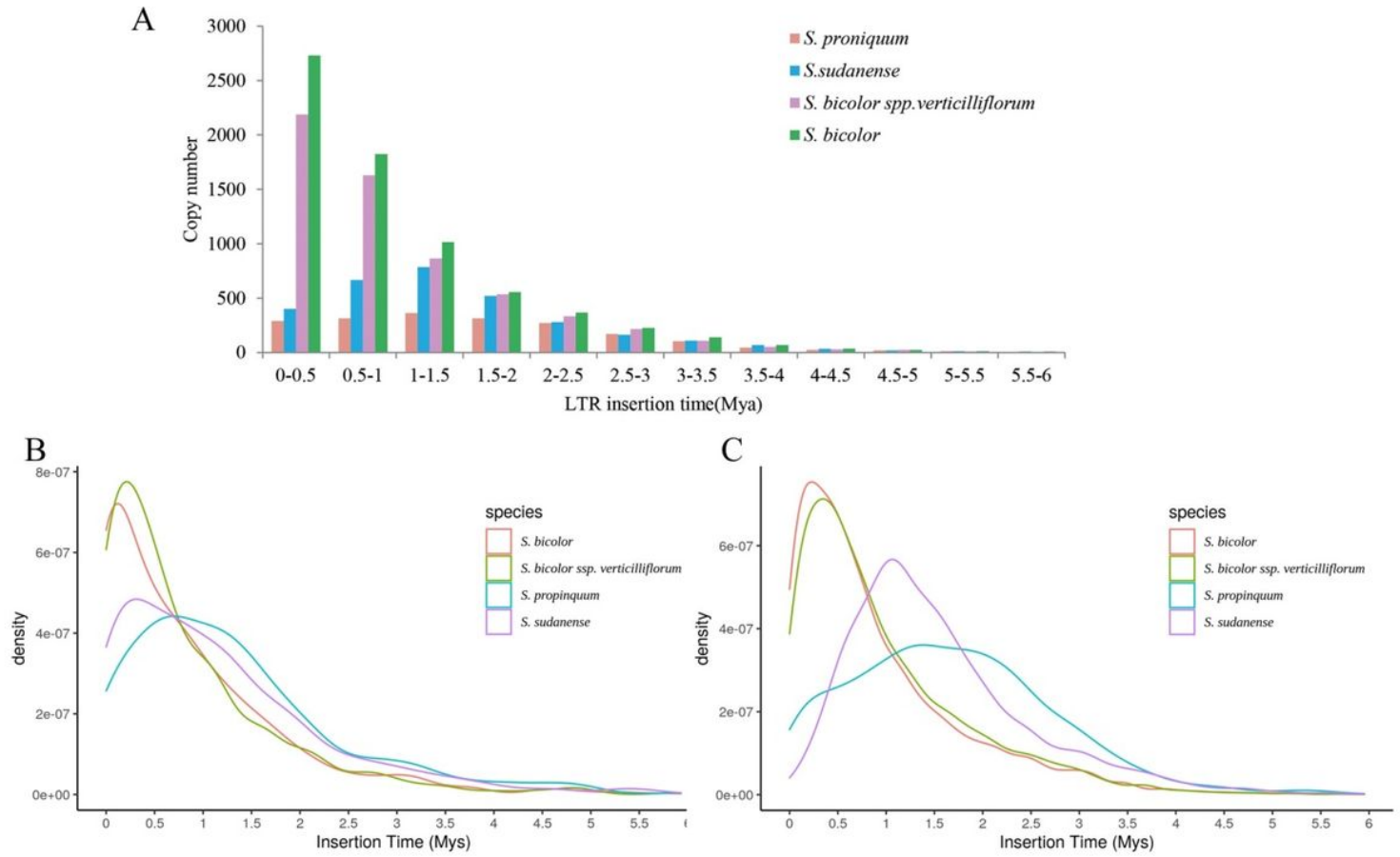

Figure 4

LTR insertion analysis. A. LTR insertion times of the four sorghum species. B. The density graph of LTR insertion times of the four sorghum species for *copia* superfamily. C. The density graph of LTR insertion times of 4 sorghum species for *gypsy* superfamily

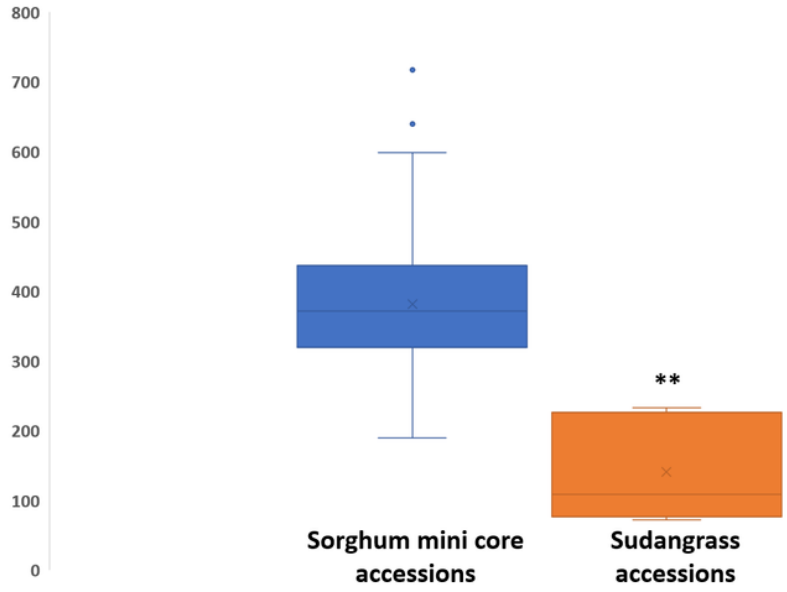

Figure 5

Higher HCN-p (ppm) in the sorghum mini core accessions than in the seven sudangrasses. ** indicates significant difference at $p < 0.01$ ($p = 0.0000383$). X inside each box in the boxplot represents the mean and horizontal line the median value of each data group.

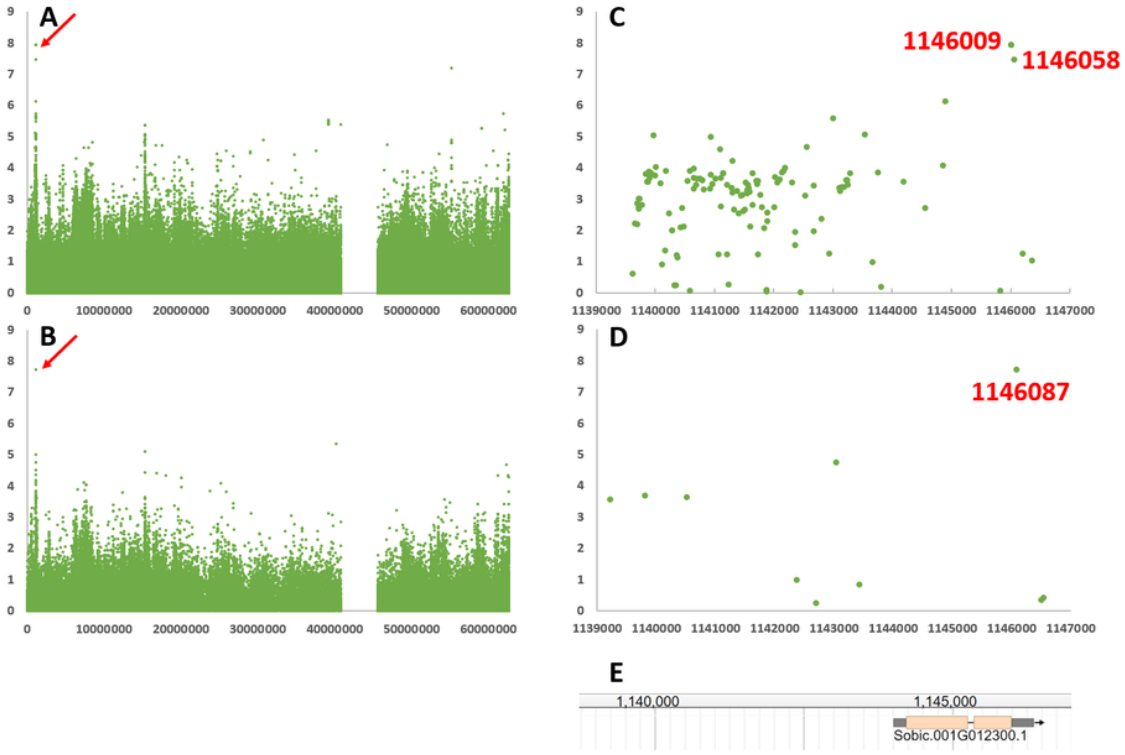

Figure 6

A QTL on sorghum chromosome 1 (indicated by red arrow in A and B) associated with HCN-p. **A.** Manhattan plot of SNPs on chromosome 1. **B.** Manhattan plot of indels on chromosome 1. **C.** The peak region indicated by red arrow in **A** magnified. **D.** The peak region indicated by red arrow in **B** magnified. **E.** Genomic region from **C** and **D** showing all annotated genes. For **A**, **B**, **C**, and **D**, the x axis is physical distance in bp and the y axis is -log(p).

Supplementary Files

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

- SupplementaryFig.1.jpg
- SupplementaryFig.2.png
- SupplementaryFig.3.png
- SupplementaryFig.4.png
- SupplementaryFig.5.png
- SupplementaryFig.6.png
- SupplementaryTable1.docx
- SupplementaryTable2.docx
- SupplementaryTable3.docx

- SupplementaryTable4.docx
- SupplementaryTable5.xlsx
- SupplementaryTable6.docx
- SupplementaryTable7.docx
- SupplementaryTable8.csv
- SupplementaryTable9.csv
- SupplementaryTable10.csv