NAR Genomics and Bioinformatics, 2024, **6**, lqae097
https://doi.org/10.1093/nargab/lqae097
Advance access publication date: 9 August 2024
High Throughput Sequencing Methods

High-quality chromosome scale genome assemblies of two important Sorghum inbred lines, Tx2783 and RTx436

Bo Wang[1,†], Kapeel Chougule[1,†], Yinping Jiao[1,2], Andrew Olson[1], Vivek Kumar[1], Nicholas Gladman[1,3], Jian Huang[4], Victor Llaca[5], Kevin Fengler[5], Xuehong Wei[1], Liya Wang[1], Xiaofei Wang[1], Michael Regulski[1], Jorg Drenkow[1], Thomas Gingeras[1], Chad Hayes[6], J. Scott Armstrong[7], Yinghua Huang[8,9], Zhanguo Xin[6] and Doreen Ware[1,3,*]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
[2]Texas Tech University, 1006 Canton Ave, Lubbock, TX 79409-2122, USA
[3]USDA ARS Robert W. Holley Center for Agriculture and Health Cornell University, Ithaca, NY, USA
[4]Department of Plant and Soil Sciences, Oklahoma State University, Stillwater, OK 74078-6028, USA
[5]Corteva Agriscience™, 8325 NW 62nd Avenue, Johnston, IA 50131, USA
[6]U.S. Department of Agriculture-Agricultural Research Service, Plant Stress and Germplasm Development Unit, Cropping Systems Research Laboratory, Lubbock, TX 79415, USA
[7]Peanut and Small Grains Research Unit, 1301 N. Western Rd. Stillwater, OK 74075, USA
[8]USDA-ARS Plant Science Research Laboratory, 1301 N. Western Road, Stillwater, OK 74075-2714, USA
[9]Dept. of Plant Biology, Ecology, and Evolution, 301 Physical Sciences, Stillwater, OK 74078-3013, USA

[*]To whom correspondence should be addressed. Tel: +1 516 367 6979; Fax: +1 516 367 6851; Email: ware@cshl.edu
[†]The first two authors should be regarded as Joint First Authors.

Abstract

Sorghum bicolor (L.) Moench is a significant grass crop globally, known for its genetic diversity. High quality genome sequences are needed to capture the diversity. We constructed high-quality, chromosome-level genome assemblies for two vital sorghum inbred lines, Tx2783 and RTx436. Through advanced single-molecule techniques, long-read sequencing and optical maps, we improved average sequence continuity 19-fold and 11-fold higher compared to existing Btx623 v3.0 reference genome and obtained 19 and 18 scaffolds (N50 of 25.6 and 14.4) for Tx2783 and RTx436, respectively. Our gene annotation efforts resulted in 29 612 protein-coding genes for the Tx2783 genome and 29 265 protein-coding genes for the RTx436 genome. Comparative analyses with 26 plant genomes which included 18 sorghum genomes and 8 outgroup species identified around 31 210 protein-coding gene families, with about 13 956 specific to sorghum. Using representative models from gene trees across the 18 sorghum genomes, a total of 72 579 pan-genes were identified, with 14% core, 60% softcore and 26% shell genes. We identified 99 genes in Tx2783 and 107 genes in RTx436 that showed functional enrichment specifically in binding and metabolic processes, as revealed by the GO enrichment Pearson Chi-Square test. We detected 36 potential large inversions in the comparison between the BTx623 Bionano map and the BTx623 v3.1 reference sequence. Strikingly, these inversions were notably absent when comparing Tx2783 or RTx436 with the BTx623 Bionano map. These inversion were mostly in the pericentromeric region which is known to have low complexity regions and harder to assemble and suggests the presence of potential artifacts in the public BTx623 reference assembly. Furthermore, in comparison to Tx2783, RTx436 exhibited 324 883 additional Single Nucleotide Polymorphisms (SNPs) and 16 506 more Insertions/Deletions (INDELs) when using BTx623 as the reference genome. We also characterized approximately 348 nucleotide-binding leucine-rich repeat (NLR) disease resistance genes in the two genomes. These high-quality genomes serve as valuable resources for discovering agronomic traits and structural variation studies.

Introduction

Sorghum bicolor (L.) Moench, the fifth most economically important cereal crop in the world after maize, rice, wheat and barley (1), is known as the 'camel of the grass family' due to its high heat and drought tolerance. Hence, accelerating crop improvement in sorghum is key to ensuring global food and energy security in the context of climate change (2). Moreover, its small and compact genome relative to other C4 grasses makes it an excellent model for genomic studies (3). Previous population genomic and genome-wide association studies of agroclimatic traits in sorghum provided a basis for crop improvement through marker-assisted breeding and genomic selection (2). In addition, whole-genome sequencing of different sorghum lines spanning a wide range of geographic origins indicated that sorghum offers underdeveloped genetic resources that are unique among the major cereals (3,4).

Here, we report the development of two new high-quality chromosome-level reference assemblies of sorghum: Tx2783, a widely utilized pollinator parent with sugarcane aphid resistance, and RTx436, another widely utilized parental inbred restorer line with known general combinability but known to be sugarcane aphid susceptible (5). We demonstrate the utility of optical maps in identifying structural variations and correcting complex regions in genome assembly. Potential large structural variants (SVs) and SNP,INDEL variations were identified in these two new reference genomes and other publicly available accessions.

Additionally, we performed comparative analysis which encompassed 26 plant genomes, comprising 18 sorghum

Received: September 12, 2023. **Revised:** July 1, 2024. **Editorial Decision:** July 17, 2024. **Accepted:** July 23, 2024
Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics 2024.

genomes and 8 outgroup species, to construct gene trees utilizing protein-coding genes. From these gene trees, representative models were selected across the 18 sorghum genomes to establish a sorghum pan-gene index, delineating core, softcore, and shell genes. Furthermore, we characterized highly evolving nucleotide-binding and leucine-rich repeat (NLR) genes and their associated integrated domains across all sorghum lines. These findings offer valuable insights into the genetic diversity of sorghum and provide potential resources for breeding sorghum varieties with enhanced resistance to sugarcane aphids.

Materials and methods

Bionano map

Ultra-high molecular weight nuclear DNA (uHMW nDNA) was isolated using a modified version of the Bionano Plant Tissue DNA Isolation Base Protocol (https://bionanogenomics.com/wp-content/uploads/2017/01/30068-Bionano-Prep-Plant-Tissue-DNA-Isolation-Protocol.pdf). Approximately 0.7–2 g of healthy young leaf tissue was collected from seedlings two weeks after germination. Leaf tissue was treated in a 2% formaldehyde fixing solution, washed, diced, and homogenized using a Qiagen TissueRuptor probe. The resultant homogenate was filtered iteratively through 100- and 40-μm cell strainers, and then pelleted by centrifugation at 2500 × *g* for 20 min. Free nuclei were concentrated by step gradient centrifugation and pelleted by standard centrifugation at 2500 × *g* for 10 min. The nuclei pellet was embedded into a low-melting-point agarose plug, followed by treatment with proteinase K and RNase A. Agarose plugs were washed four times in Bionano wash Buffer and five times in Tris–EDTA buffer, pH 8.0. Purified DNA was recovered by digesting the plug with agarase followed by drop dialysis against TE pH 8.0.

Direct label and stain (DLS) was used in combination with a Bionano Saphyr system to generate chromosome-level optical maps. Bionano DLS uses the DLE1 enzyme, which labels DNA molecules by recognizing CTTAAG sites and attaching a single fluorophore. DLS was performed using the Bionano Direct Label and Stain Kit (https://bionanogenomics.com/wp-content/uploads/2018/04/30206-Bionano-Prep-Direct-Label-and-Stain-DLS-Protocol.pdf) with a few modifications. Approximately 1 μg of sorghum uHMW nDNA was mixed with DLE-1 Enzyme, DL-Green label and DLE-1 Buffer, and then incubated for 2:20 h at 37°C. The reaction was stopped by incubation for 20 min at 70°C, followed by digestion with proteinase K for 1 h at 50°C. Unincorporated DL-Green label was removed from the reaction by absorption onto a nitrocellulose membrane. The labeled, cleaned-up DNA sample was combined with Flow Buffer, DTT, incubated overnight at 4°C, and quantified. The DNA backbone was stained by the addition of Bionano DNA Stain at a final concentration (0.11 μg/μl) of final DNA. Finally, the labeled, cleaned-up, and stained sample was loaded onto a single Bionano chip flow cell. DNA molecules were electrophoretically separated, stretched, imaged, and digitized using the Bionano Genomics Saphyr System and server (https://bionanogenomics.com/wp-content/uploads/2017/10/30143-Saphyr-System-User-Guide.pdf).

PacBio and 10× chromium sequencing collection

DNA was isolated from approximately 50 g of fast-frozen young leaf tissue using a protocol described by (6). Leaf material was ground in liquid nitrogen and incubated for 15 minutes in cold NIB buffer (10 mM Tris–HCl pH 8.0. 10 mM EDTA pH 8.0, 100 mM KCl, 500 mM sucrose, 4 mM spermidine trihydrochloride, 1 mM spermine tetrahydrochloride) containing 0.1% 2-mercaptoethanol. The resultant homogenate was filtered twice through Miracloth and mixed with 5% NIBT (NIB + 10% Triton X-100). Free nuclei and cell debris were concentrated by centrifugation at 2500 × *g* for 15 minutes and washed with NIB + 0.1% 2-mercaptoethanol followed by a second centrifugation at 2500 × *g* for 15 min. The pellet was embedded in LMP agarose plugs, treated with proteinase K, melted, and digested with agarase. High-speed (30 000 × *g*) centrifugation cycles were performed to concentrate and remove solids and recover the DNA in the supernatant. PacBio library preparation was performed using the Pacific Biosciences SMRTbell Template Prep Kit 1.0 following the protocol for >30-kb libraries (https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Preparing-HiFi-SMRTbell-Libraries-using-SMRTbell-Template-Prep-Kit-1.0.pdf) with some modifications.

Prior to library construction, the DNA was sheared to approximately 50 kb using a BioRuptor (Diagenode) and pre-sized using the >20-kb high-pass broad-range protocol on a PippinHT (Sage Science). A second size selection for the final library was performed using the >20-kb broad-range PippinHT protocol, followed by a final DNA repair reaction. The quality of DNA at different steps of the process was evaluated using a FEMTO Pulse automated pulsed-field analyzer (Agilent Technologies, Wilmington, DE, USA). Quantification was performed using the Qubit dsDNA BR and HS reagents Assay kits on a Qubit 3.0 system. Sequencing was performed on a PacBio Sequel v6.0 system using 1M v3 SMRT Cells with 3.0 chemistry and diffusion loading. In total, six SMRT Cells were used for Tx2783 and RTx436. 10X Chromium was performed according to standard protocols using a non-sheared aliquot of the DNA used for PacBio library construction. Approximately 1 ng of DNA was loaded with 10X Chromium reagents and gel beads onto a Chromium chip (https://support.10xgenomics.com/de-novo-assembly/library-prep/doc/user-guide-chromium-genome-reagent-kit-v1-chemistry). The Chromium libraries were sequenced on an Illumina HiSeq2500 system. For Tx2783, a total of 231 083 110 clusters were obtained, corresponding to 69.8 Mb or 87× genome coverage. For RTx436, a total of 229 285 050 clusters were obtained, corresponding to 69.3 Mb or 86× genome coverage.

Long- and short-read sequencing

Long-read data were generated using the Pacific BioSciences (Menlo Park, CA, USA) Sequel platform. Six SMRT cells were sequenced for both samples with 10-hr movies and v6 chemistry. Raw subreads were filtered to a 3-kb minimum for both samples, generating 75× and 64× coverage depth for Tx2783 and RTx436, respectively. The raw subread N50 lengths were 23.1 kb (Tx2783) and 24.7 kb (RTx436). Linked short-read data were generated by sequencing of 10× Genomics (Pleasanton, CA, USA) Chromium on the HiSeq 2500 plat-

form (Illumina, San Diego, California). The coverage depths and mean molecule lengths for the Tx2783 and RTx436 Chromium libraries were 76×/62.9 kb and 84×/98.5 kb, respectively.

Genome assembly and polishing

Raw PacBio subreads were corrected and assembled using Canu v1.8 (7) with the following parameters varying from the defaults: 'correctedErrorRate = 0.065 corMhapSensitivity = normal ovlMerDistinct = 0.99'. The resultant contigs were filtered to a minimum contig length of 30 kb. Beyond the sequence consensus process that Canu performs after assembly, additional sequence polishing was performed by aligning raw PacBio subreads to the contig assembly using pbmm2 v0.12.0 and applying the Arrow algorithm from the Genomic Consensus package (v2.3.2) to get consensus calls. Both of these tools were obtained from pbbioconda (https://github.com/PacificBiosciences/pbbioconda). To further increase the consensus sequence accuracy, the long read contig assembly was complemented with Chromium linked short-read 'clouds', which have higher unique mappability than standard Illumina paired-end libraries. Chromium datasets were aligned to sequence contigs with Long Ranger v2.2.2.

The assembly improvement tool Pilon v1.22 (https://github.com/broadinstitute/pilon) was used to correct individual base errors and small indels from the consensus of Chromium data aligned to the contigs using the parameter '–fix bases –minmq 30'. To decrease the compute time, a separate Pilon job was run for each contig against Chromium alignments specific to that contig. Contig-specific alignments were created from the Long Ranger output BAM file using samtools v1.9 (https://github.com/samtools/samtools).

Genome mapping

Genome maps for Tx2783 and RTx436 were generated on the Bionano Genomics (San Diego, California) Saphyr platform using the Direct Label and Stain (DLS) system. For Tx2783, DLE-1–labeled molecule data from two flow cells (one chip) were filtered to create a data subset with a molecule N50 of 682 kb and 189 × coverage. The filtered molecule dataset for Tx2783 was assembled via the Bionano Genomics Access software platform (Solve3.2.2_08 222 018) with the configuration file optArguments_nonhaplotype_noES_noCut_DLE1_saphyr.xml. The map assembly for Tx2783 consisted of 27 genome maps with a genome map N50 of 36.9 Mb and a total map length of 732.1 Mb.

For RTx436, DLE-1 labeled molecule data from two flow cells (1 chip) was filtered to create a dataset with a molecule N50 of 441 kb and 267× coverage. The filtered molecule dataset for RTx436 was assembled as described above, generating a map assembly of 28 genome maps with a genome map N50 of 37.7 Mb and a total map length of 723.6 Mb.

Hybrid scaffolding

An initial hybrid scaffolding was generated from the polished contigs and the Bionano genome maps using the Bionano Genomics Access software (Solve3.3_10 252 018) and the DLE-1 configuration file hybridScaffold_DLE1_config.xml to identify potentially problematic (smaller maps nearly identical to larger maps, low-coverage) or non-contributing genome maps. After this assessment, 19 genome maps from Tx2783 (min length = 8.6 Mb) and 18 genome maps for RTx436 (min length = 8.1 Mb) were selected to generate hybrid scaffolds.

As part of the hybrid scaffolding process, chimeric misassemblies in the contigs were identified and cut to resolve conflicts relative to the genome maps. In addition to autoconflict resolution, manual curation was performed to resolve overlapping contigs that were not addressed by the hybrid scaffolding workflow. A list of contig pairs that overlap in map space was generated from the gap file in the hybrid scaffold output directory, yielding 58 (Tx2783) and 51 (RTx436) overlapping pairs. Additional contig cuts were added to the conflict resolution file to best resolve these issues. Embedded contigs are another assembly issue not resolved by the hybrid scaffolding workflow. A list of contig pairs in which a smaller contig is embedded within a collapsed region of a larger contig was also generated from the gap file in the hybrid scaffold output directory, yielding three (Tx2783) and eight (RTx436) embedded contig pairs. Similarly, additional conflict cuts were made in the larger contig to allow incorporation of the smaller contig after the hybrid scaffolding was re-run. In the final assembly, Tx2783 had 19 hybrid scaffolds (scaffold N50 = 36.0 Mb, total scaffold length = 696.8 Mb) with 310 unscaffolded contigs with a total length of 27.1 Mb. RTx436 had 18 hybrid scaffolds (scaffold N50 = 37.6 Mb, total scaffold length = 697.2 Mb) with 325 leftover contigs that were not scaffolded with a combined length of 25.0 Mb).

Creating pseudomolecules

With an average of 1.85 hybrid scaffolds per chromosome for the two lines, it is straightforward to create chromosome-scale pseudomolecules using the *Sorghum bicolor* v3.1 reference genome as a guide (https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Sbicolor). A generalized approach was used that rapidly maps scaffolds to a reference genome and determines their relative position and orientation. First, each scaffold was chunked into 100-bp fragments and then aligned to the *Sorghum bicolor* v3.1 reference genome using minimap2 v2.10(8). Then, a custom script was used to determine the chromosome position and orientation from the alignment of scaffold 'read clouds'. All scaffolds were placed with this method. The remaining unscaffolded contigs were concatenated with 100-bp N-gaps and placed into Chr00.

Integrating plastid genomes into the assembly

It can be challenging to assemble plastid genomes from whole-genome PacBio libraries and distinguish them from plastid DNA integrated into the chromosome. An alternative approach is to insert a copy of an existing plastid genome reference sequence and remove incomplete plastid assemblies from Chr00. A copy of the *Sorghum bicolor* chloroplast complete genome (GenBank: NC_008602.1) and *Sorghum bicolor* mitochondrion complete genome (GenBank: NC_008360.1) were added to the assembly as ChrC and ChrM, respectively. Only Tx2783 exhibited a SNP or small indel when the Canu-corrected reads were aligned with minimap2 and checked for variants with Pilon, which had a 1-bp deletion relative to the mitochondria reference. A modified version of ChrM incorporating this deletion was added to the Tx2783 assembly. Incomplete plastid sequence contigs in Chr00 were identified by BLASTN (-F f -e 1e-200 -W32). Contigs with greater than 99% identity over more than 20 kb relative to the *Sorghum bicolor* plastid reference genomes were removed from the

assembly. Manual curation of both assemblies was performed to fill gaps and better resolve the chromosomal chloroplast insertion on Chr09.

Annotation of the Tx2783 and RTx436 assembly

Gene annotations were performed using a strategy that combined evidence-based and *ab initio* gene predictions (9). Genome-guided transcript assemblies were performed using transcriptome data from seven tissue types across the juvenile, vegetative, and reproductive stages of development (Supplementary Data S1) using four different assemblers: Trinity (v2.8.4) (10), StringTie (v1.3.5) (11), Cufflinks (v2.1.1) (12) and PsiCLASS (v1.0.1) (13). Representative sets of transcripts were identified and annotated as genes using Mikado (v2.0rc6) (14). High-quality RNA-seq reads from each library were mapped to Tx2783 and RTx436 genomes indexed by STAR (v2.7.0d) (15) using a two-pass mapping approach and default settings. The SAM output from individually mapped RNA-seq libraries was then pooled, sorted, and indexed for transcript assembly programs using Picard (v2.7.0) (http://broadinstitute.github.io/picard). The transcript assemblers were run using default options, except for Trinity, for which the maximum intron size was set to 10000. The transcript assembly fasta file from the Trinity output was converted to GFF3 by aligning the transcripts to indexed genomes with GMAP (v2019-03-15) (-f gff3_match_cdna) (16). Portcullis (v1.1.2) (17) was used to generate high-quality splice junctions from the merged mapped reads.

Preliminary transcripts were refined for Mikado by (i) merging all transcripts and removing the redundant copies, (ii) processing using TransDecoder (v5.5.0) (10) (to identify open reading frames) and (iii) aligning with BLASTX (v2.2.29+) (18) against SwissProt Viridiplantae proteins (to identify full-length transcripts). Default options were used for TransDecoder. For BLASTX, maximum target sequences were set to 5 and output format to xml. Inputs for Mikado included all transcript assemblies, Portcullis-generated splice sites, and a plants.yaml scoring matrix. The output GFF3 file was used to extract transcripts and proteins using the gffread utility from the Cufflinks package.

Further structural improvements to Mikado-generated transcripts were completed using the PASA (v2.3.3) (19) genome annotation tool. The inputs for PASA included 209 835 maize ESTs derived from GenBank, Mikado-assembled transcripts for Tx2783 and RTx436, 37 655 sorghum iso-seq transcripts from 11 developmental tissues (20) and 31 881 sorghum full-length cDNAs from RIKEN that were filtered for intron retention (21). PASA was run with default options; in the first step, transcript evidence was aligned to the masked sorghum genomes using GMAP and Blat (v.36) (22). The full-length cDNA and Iso-seq transcript IDs were passed in a text file (-f FL.acc.list) during the PASA alignment step. PASA updated the models, providing UTR extensions, as well as novel and additional alternative isoforms. PASA-generated models were passed through the MAKER-P (v3.0) (23) annotation pipeline as model_gff along with all the transcript and protein sequences to yield Annotation Edit Distance (AED) scores (24) to assess the quality of annotations.

Transposon element (TE)-related genes were filtered using the TEsorter tool (25), which uses the REXdb (viridiplantae_v3.0 + metazoa_v3) (26) database of TEs. To supplement the evidence annotation, the protein sequences from Mikado transcripts and RNA-seq data were used for *ab initio* gene model prediction with BRAKER (27). Non-overlapping BRAKER gene models were updated with PASA, filtered for TE-related genes, provided with AED scores using MAKER-P, and added to the evidence set. We further filtered this combined set with the criterion AED <0.75, and then applied phylogeny filters by aligning protein sequences to maize, rice, *Brachypodium*, and Arabidopsis proteins to identify conserved and lineage-specific genes. The genes were loaded in Ensembl core databases and quality-checked to identify transcripts with incomplete CDSs, which were programmatically corrected and also checked for translation errors. Transcripts with complete CDSs were tagged as protein-coding and those with incomplete CDSs as non-coding.

The determination of protein-coding and non-coding genes was based on two main criteria: AED score, which ranges from 0 to 1, and conservation versus specificity as determined by protein alignments using USEARCH (28). Initially, we consolidated evidence-based gene models that did not overlap, creating a comprehensive set of gene models. This combined set was then refined using an AED threshold of <0.75. Subsequently, phylogenetic filters were applied by aligning protein sequences to those of Maize (29), Rice (30), Brachypodium (31), and Arabidopsis (32) to identify genes that were conserved across species or specific to certain lineages. Genes that lacked hits to any of the outgroup species were categorized as 'specific'. The filtered and classified genes were uploaded into Ensembl core databases and assigned corresponding biotypes.

Lineage-specific genes, many of which were single-exon genes (~50%), were further examined to determine if their transcripts contained complete coding sequences (CDS), defined by the presence of a start codon (ATG) and a stop codon (TAA, TAG or TGA). Lineage-specific genes meeting these criteria were included in the conserved gene set, while those lacking complete CDS were classified as non-coding. Detailed statistics on these annotation features are provided in (Supplementary Data S2). Functional domain identification was completed with InterProScan (v5.38–76.0) (33). TRaCE (34) was used to assign canonical transcripts based on domain coverage, protein length, and similarity to transcripts assembled by Stringtie. Finally, the protein coding annotations were imported to Ensembl core databases, verified, and validated for translation using the Ensembl API (35).

Rampage library construction, data generation and analysis

This protocol is a modified version of a previously published method (36). Prior to incubation with Terminator™ 5′-Phosphate-Dependent Exonuclease (TEX) (Lucigen) to remove all residual RNAs containing 5′ monophosphate, we removed all ribosomal RNAs using the RiboMinus™ Plant Kit for RNA-Seq (Thermo Fisher Scientific).

We then performed first-strand synthesis using the SMARTer Stranded Total RNA Kit V2- Pico Input Mammalian (Takara). Following purification with RNAclean XP (Beckman Coulter), 5′ cap oxidation, 5′ cap biotinylation, RNase I digestion, and streptavidin pulldown (Cap Trapping) were performed as described in (36).

Amplification of purified cDNAs (two rounds of PCR to attach Illumina adapters and amplify the libraries) followed by AMPure XP cleanup (Beckman Coulter) was done using the

'SMARTer Stranded Total RNA Kit v2' (Takara) according to protocol.

All samples were processed separately, quantitated on a 2100 Bioanalyzer using a HS-DNA-Chip (Agilent), and adjusted to a concentration of 10 nM. Libraries were then pooled at equimolar concentration and sequenced on an Illumina NextSeq 550 Sequencer. RAMPAGE reads were aligned to the reference using STAR 2.7 on SciApps (37). The mapped reads from each tissue were clustered using Paraclu (38). The BAM alignment files from STAR were converted to BED using the bamtobed tool and grouped using the groupBy tool from BEDTools v2.29.2 (39) to sum up reads that started at the same position and on the same strand as the input to Paraclu. Paraclu was run with default settings with the minimal number of reads to form a cluster; -minValue was set to 10. The paraclu-cut.sh script within Paraclu was used to simplify and remove clusters with length > 200, max density/min density <2, or that were contained within another cluster. Peaks with the clusters were identified using scripts provided in this Github repository : https://github.com/davetang/paraclu_prep. TSS profile plots for scores over the TSS region using an annotated 5′ UTR were generated using deeptools2 (40).

Phylogenetic gene trees

Genome cores serve as the fundamental basis for constructing protein-based gene trees. In the fourth release of SorghumBase (41) (https://www.sorghumbase.org/), we incorporated 18 sorghum genomes, including (TX2783 and RTx436), and diverse outgroup species (Zea mays, Oryza sativa, Vitis vinifera, Arabidopsis thaliana, Selaginella moellendorffii, Populus trichocarpa, Drosophila melanogaster and Chlamydomonas reinhardtii). These genomes were utilized as inputs for generating protein-based gene trees through Ensembl Protein Comparative phylogenetic analysis (42). The resulting analyses produced 31 210 protein-coding gene family trees, constructed by considering the peptides encoded by the canonical transcript of each of the 829 431 individual genes (870 922 input proteins) from the 26 genomes. These gene trees offer a structural framework for the phylogenomic dating of sorghum genes, enabling the identification of orthologs and paralogs. This framework facilitates the exploration of genetic relationships both between and within species, contributing to the comprehensive characterization of the species pan-gene set.

Pan-gene analysis

Utilizing the gene trees, representative pan-gene models were selected from a sorted list of 18 Sorghum genomes including (TX2783 & RTx436). The pan-gene protein was further classified by taxonomic age which was determined by orthology. Specifically, for each pan-gene, we selected a representative species, such as Arabidopsis thaliana for Viridiplanteae, Oryza sativa for Poaceae, and Zea mays for Andropogoneae. If the protein encoded by the pan-gene was orthologous to any of these representative species in the protein coding gene tree, it was categorized accordingly. Otherwise, it was labeled as Sorghum specific. Additionally, the pan-genes were further categorized as core if they were identified as orthologs in all 18 sorghum accessions, soft-core if present in any 2–17 accessions, or cloud if found exclusively in one accession.

Identification of disease resistance genes-NLR and NLR-ID

NLR, denoting Nucleotide-binding domain and Leucine-rich repeat, signifies a vital protein class involved in plant disease resistance (43). In contrast, NLR-ID, standing for NLR-Integrated Domain (ID), describes NLR proteins with added integrated domains beyond the standard nucleotide-binding site (NBS) domain and leucine-rich repeats (LRR) (44). NLR and NLR-ID's Integrated Domains (IDs) were characterized in Sorghum annotations using the plant_rgenes pipeline (https://github.com/krasileva-group/plant_rgenes) (45). NLR proteins are identified based on the presence of Nucleotide-Binding Adaptor Shared by APAF-1, R proteins and CED-4(NB-ARC) domain (IPR002182) (46) while NLR-IDs are identified based on the presence of non-NBS, non-LRR domains (e-value cutoff 1e-3). The number of NB-ARC containing proteins was plotted using R package ggplot2 (47). The NB-ARC domain alignment was manually curated for the presence of NB-ARC domain functional motifs, including Walker A, WALKER-B, RNBS-C, GLPL and RNBS-D. The NLR phylogeny was determined using RAxML MPI (v8.2.9,-f a, -x 12 345, -p 12 345, -# 100, -m PROTCATJTT). The phylogeny was visualized and re-rooted on the longest internal branch in Interactive Tree of Life (iTOL) (48). In addition, Illumina raw reads published study (49) were aligned to the Tx2783 reference genome using STAR (15) with a minimum intron length set to 20 bp and a maximum intron length set to 50 kb, with default settings for other parameters. Quantification of genes and isoforms was performed using cufflinks version 2.2.1 (12). A k-means clustering using R Bioconductor package 'Mfuzz' (50) was done to cluster dynamically expressed genes based on their expression profiles across different time points after sugarcane aphid infestation.

Results

Chromosome-level genome assembly of two sorghum inbred lines

To characterize genetic variation, and disease resistance genes in the sorghum population and to provide support for modern breeding, we selected two important sorghum inbred lines for genome construction, Tx2783 and RTx436. Line Tx2783 (PI 656001) was originally bred for resistance to sorghum greenbug (*Schizaphis graminum*) and also exhibits high sugarcane aphid resistance (49). The other line, RTx436 (PI 561071), is a widely adapted pollinator parent used in the development of high-yielding hybrids. RTx436 is also commonly used in the development of traditional food-grade sorghum hybrids with white grain color and tan glumes, which are suitable traits for many food processors. These two sorghum lines, Tx2783 and RTx436, were sequenced using PacBio CLR technology to coverage of 76× (reads N50 = 23.1 kb) and 61× (reads N50 = 24.7 kb), respectively. The assembly effort generated contigs with N50 lengths of 25.6 Mb for Tx2783 and 14.4 Mb for RTx436. In addition, we generated Bionano molecules for Tx2783 and RTx436 that yielded genome maps of 721.504 and 723.680 Mb, respectively, with N50 lengths of 36.987 and 37.781 Mb (Supplementary Table S1).

The chromosomes of these two genomes were constructed using hybrid scaffolds generated from Bionano genome

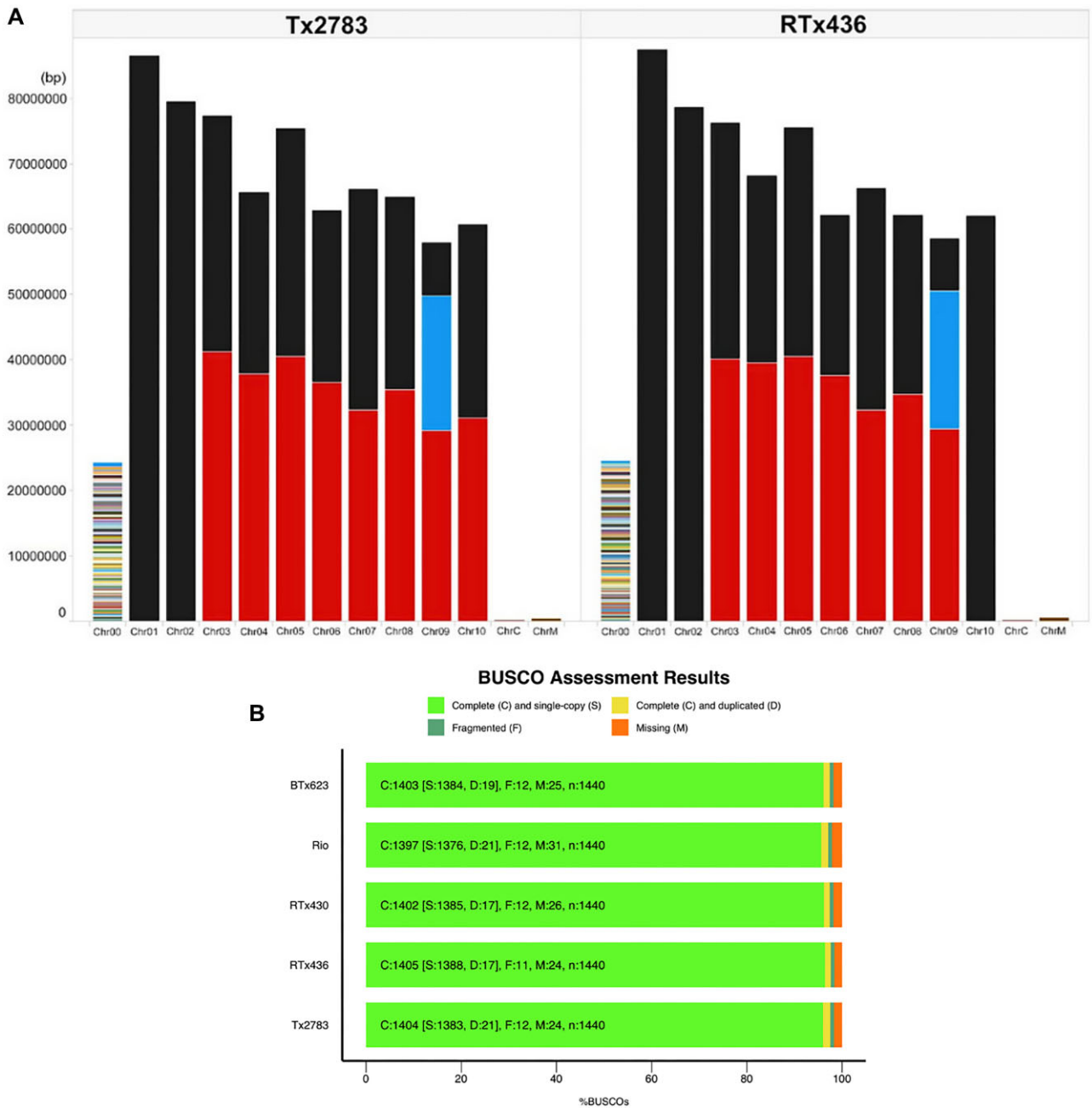

Figure 1. Number of scaffolds per chromosome and BUSCO assessment. (**A**) Number of scaffolds per chromosome for the Tx2783 and RTx436 genome assemblies. Each colored bar represents a scaffold. Most scaffold breaks occur at the centromeres. Chr00: Contigs unplaced into the chromosome. ChrC: Chloroplast sequence. ChrM: mitochondria sequence. (**B**) BUSCO assessment of the indicated sorghum genomes.

maps. Most of the chromosomes consisted of two scaffolds (Figure 1A, Table 1). These two genomes were of higher quality than the published sorghum reference genomes of BTx623 (51) with 19-fold and 11-fold higher average sequence continuity, respectively, and a total assembly size that is much closer to the estimated size of the sorghum genome, ~730 Mb. In addition, the assembly of repeat space in the two genomes using the LTR assembly index (52) was comparable to the reference BTx623 genome (Supplementary Table S2). Gene space assessment using BUSCO (53) confirmed the completeness of the Tx2783 and RTx436 genome assemblies with scores above 96% for single-copy orthologs derived from embryophyta lineage (Figure 1B).

Transposable element and gene annotation of the Tx2783 and RTx436 genomes

Transposon elements (TEs) were annotated using Extensive de-novo TE Annotator (EDTA)(54) for the Tx2783, RTx436, and BTx623 genomes. The annotation revealed that 68.97%, 69.34% and 69.35% of the genome sequences were annotated as TEs in Tx2783, RTx436, and BTx623, respectively. The majority (53.23%, 53.29% and 53.19%) of TEs in each genome were annotated as retrotransposons. Of those, LTR-Gypsy was the most abundant, representing 36.52%, 37.20% and 37.98% of TEs in the Tx2783, RTx436, and BTx623 genomes, respectively (Supplementary Table S3). In addition, we also identified 10 984 (79.5%), 11 207

Table 1. Genome assembly statistics for Tx2783 and RTx436

	Btx623	Tx2783	RTx436
Sequencing coverage	8×	76×	61×
No. of contigs	2688	447	464
Contig N50 (Mb)	1.3	25.6	14.4
Total contig length (Mb)	675.4	711.9	706.5
No. of genome map	37	19	19
Genome map N50 (Mb)	35.1	37.0	37.8
Total genome map length (Mb)	721.3	707.8	712.6
No. of hybrid scaffold	N/A[a]	19	18
Hybrid Scaffold N50 (Mb)	N/A[a]	36.2	37.6
Total assembly length (Mb)	708.7	723.5	724.8

[a]Note: For Btx623, we have N/A for hybrid scaffold since we did not pursue hybrid scaffold assembly.

(79.7%) and 10 746 (79.2%) intact LTRs in the Tx2783, RTx436 and BTx623 genomes, respectively.

Gene calling was performed using a hybrid approach with evidence-based and *de novo* gene predictors (see Materials and Methods), and then filtered based on AED score (24) and homology to maize, *Brachypodium*, rice, or *Arabidopsis* protein sequences obtained from Gramene release 62 (55). Ultimately, this approach generated a total of 29 612 and 29 265 protein-coding genes and 4205 and 3478 non-coding genes in the Tx2783 and RTx436 genomes, respectively (Supplementary Table S4). The average gene length in the RTx436 annotation was 3900 bp, slightly higher than in Tx2783 and BTx623, which had average gene lengths of 3833 and 3714 bp, respectively. This longer gene length in RTx436 may be attributed to its genes having a longer intron average length of 514 bp compared to the other two genomes.

Additionally, Tx2783 exhibited longer protein translations, with an average peptide length of 327 amino acids (aa), compared to 279 aa in RTx436 and 281 aa in BTx623. In addition, CDSs, peptides, and 5′ and 3′ UTRs were longer in Tx2783 than in RTx436. To assess the quality of the 5′ transcriptional start sites, root and shoot tissues from Tx2783 and RTx436 were collected for RAMPAGE (RNA annotation and mapping of promoters for analysis of gene expression) assays (36). This analysis identified 228 249 and 300 633 high-confidence peaks from Tx2783 and RTx436 respectively. The narrow distribution of RAMPAGE signals (Figure 2A, B) over annotated loci confirmed our very high specificity for true TSSs in both root and shoot tissues from the Tx2783 and RTx436 genomes. The high-confidence peaks overlapped in 27 135 and 26 432 genes in Tx2783 and RTx436 respectively, indicating that the two genomes were well annotated.

Comparative analysis reveals unique protein coding genes in sorghum varieties Tx2783 and RTx436

In SorghumBase release 4, we employed 18 sorghum genomes, including TX2783 and RTx436, alongside diverse outgroup species (Zea mays, Oryza sativa, Vitis vinifera, Arabidopsis thaliana, Selaginella Moellendorffii, Populus trichocarpa, Drosophila melanogaster and Chlamydomonas reinhardtii), to generate protein-based gene trees using Ensembl Protein Comparative phylogenetic analysis. This process utilized 838 695 canonical proteins from 26 plant genomes, resulting in the construction of 31 210 protein-coding gene families, with approximately 13 956 representing sorghum-specific gene trees. Among these, 30 gene trees were specific to Tx2783 and RTx436, comprising 99 and 107 genes, respectively (Supplementary Data S3). Additionally, about 35% of these specific genes in Tx2783 and 15% in RTx436 were single exon genes, while only three genes in Tx2783 were identified as split genes by the compara pipeline. These genes showed functional enrichment for binding (molecular function) and metabolic process (biological process) (Figure 3A, B).

Upon closer inspection of the Interpro functional protein domains for the unique genes within the Tx2783 and RTx436 genomes, additional details about their molecular purpose was revealed. Specifically, the bulk of the unique Tx2783 genes seem to be involved in transposase/TE function, but interestingly there was also one aminotransferase-like gene (SbiRTX2783.02G158700) that has been shown to silence TE activity in other plant systems and is crucial for proper meristem and root formation (56). As for RTx436, there are several TOPLESS-like proteins (TRP), which are co-repressors involved in many aspects of plant development; TRPs often couple transcription factors to histone deacetylases to suppress gene expression. Additionally, there are also some Ubiquitin-like proteases in RTx436 that are involved in regulating the post-translational modification of proteins with ubiquitin-like proteins, such as SUMO (Supplementary Data S4), and that can repress the action of conjugated protein some of which have been shown to be involved with TPRs(57).

Diversity and distribution of pan-genes across sorghum genomes

We analyzed gene family trees across 18 Sorghum genomes sourced from SorghumBase, selecting representative pan-gene models based on sorted genome lists. These models were classified by taxonomic age depending on the presence of proteins from representative species. Specifically, if a gene family contained a protein from Arabidopsis thaliana, we classified it as Viridiplantae; if it contained a protein from Oryza sativa, we classified it as Poaceae; and if it contained a protein from Zea mays, we classified it as Andropogoneae. If the protein encoded by the pan-gene was orthologous to any of these representative species in the protein-coding gene tree, it was categorized accordingly. Otherwise, it was labeled as Sorghum specific. The classification hierarchy used was: Viridiplantae > Poaceae > Andropogoneae > Sorghum specific. Across the 18 Sorghum genomes, a total of 72 579 pan-genes were identified of which 34% belonged to Viridiplantae, 12% Poaceae, 4% Andropogoneae and the rest to Sorghum lineage (Figure 4A).

In a prior study, approximately 103K pan-genes were documented in 26 maize accessions (29), while around 56K were observed in 251 rice accessions (58). Similarly, approximately 44K pan-genes were identified across 13 sorghum accessions (59), and roughly 37K were noted in 54 Brachypodium accessions (60). A cumulative addition of genes per accession in the pan-gene set shows that 80% (58 024 out of 72 579) of the total pan-genes are captured in the first 12 sorghum accessions (from BTx623 to PI536008). Of the 72 579 pan-genes most of the genes were core or softcore pan-genes (75%) and only (25%) were categorized shell genes (Figure 4B).

We applied our pan-gene method using annotations from 26 maize (29) and 29 rice accessions (61–64) from recent published studies. A comparison of the pan-gene set for sorghum, maize and rice is shown in the (Supplementary Data S5). We find similar ratios of core, softcore and cloud genes in pan-

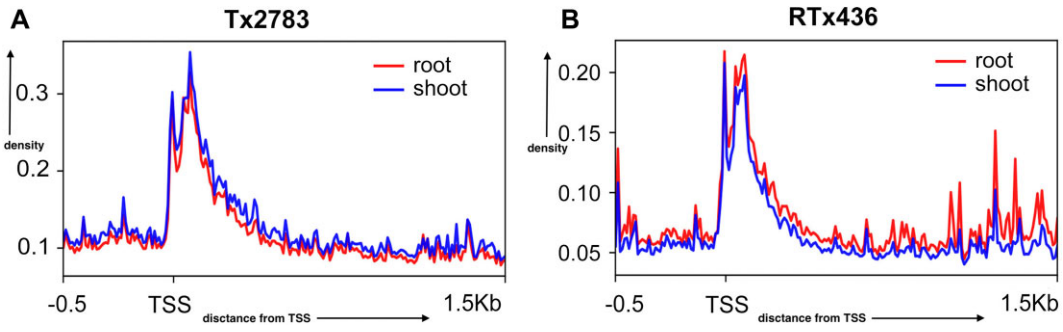

Figure 2. RAMPAGE signal enrichment around Transcription Start Site(TSS)s in Tx2783 (**A**) and RTx436 (**B**) in root and shoot tissues. The x-axis represent distance from TSS of gene in Kilo base pairs and the y-axis in a RAMPAGE TSS peak figure typically represents the read count.

gene sets for Maize and Rice. Interestingly the number of core genes in maize is almost double than in sorghum, which was expected as maize genome has undergone one ancient whole-genome duplication event (WGD)(65), whereas the number of shell or accessory genes in Sorghum is almost double that in maize but lower than in rice, indicating that the sorghum and rice pan-genomes are more open. It should also be noted that heterogeneity in annotation workflows, as well as the quality of the reference assembly will also contribute to the differences in the three different categories (66).

Disease resistance (NLR) genes in sorghum genomes

To characterize disease resistance (NLR) genes, we first assessed the distribution of NB-ARC domain in the annotated protein sequences of monocot species (*Z. mays, S. bicolor* and *B. distachyon*) and eudicot (*A. thaliana*) using the protein sequences. The results revealed that maize has far fewer R genes than sorghum (Figure 5A) with enrichment of NLR on chromosomes 5 and 8 with RTx430 having the most R genes (355) and AusTRCF317961 having the fewest (246). ∼6% of the NLRs showed the presence of integrated domains that are identified based on the presence of non-NBS non-LRR domains (Supplementary Table S5). We constructed a maximum likelihood phylogenetic tree of 333 NLR-IDs from 18 sorghum genomes. The resulting phylogeny of NLR-ID clades segregating among the sorghum lines suggests ongoing co-evolution with pathogens (67). This could be the case because NLR genes play a critical role in plant defense against pathogens, including bacteria, fungi, viruses and nematodes (68,69). As pathogens exert selective pressure on host plants, they drive the evolution of NLR genes to recognize and respond to the evolving pathogen threats (70). The segregation of NLR-ID clades among Sorghum lines indicates genetic variation in these defense genes, reflecting adaptations to different pathogen pressures in different environments (71). This ongoing co-evolutionary process helps sorghum populations maintain resistance to prevalent pathogens and adapt to changing pathogen dynamics over time. In particular, the MIC1 NLR clade which has been characterized as a fast evolving NLR clade among Poaceae (51) also showed segregation among the wild and cultivated lines (Figure 5B).

In 2013, sugarcane aphid emerged as a major insect pest of sorghum crops in North America (72). The line Tx2783 is highly resistant to sugarcane aphid, and thus has great potential to boost breeding targeted at aphid resistance (49). To investigate how gene expression is impacted by sugarcane aphid, we remapped the RNA-Seq data before and after sugarcane aphid infestation from a previous study (49) to our Tx2783 genome. We also observed diverse expression patterns for all genes throughout the genome (Figure 6A). We then grouped all genes into 16 clusters based on their expression levels (Figure 6B). This analysis revealed that 3944 genes were upregulated relative to control plants after 5, 10 and 15 days of infestation. In particular, 173 NLR genes were expressed before or after sugarcane aphid infestation, and 27 NLR genes were continuously upregulated after 5, 10 and 15 days of infestation. The overall expression of most NLR genes increased after 5 and 10 days of infestation, with a slight change at 15 days (Figure 6C).

Characterization of large structure variations using optical maps

We initially assessed large-scale structural variations between the accessions by comparing alignments between the BTx623 Bionano map and the sorghum v3.1 reference sequence. This comparison revealed several large inversions (Figure 7A), which were also found in the alignment between RTx436, Tx2783 and the v3.1 reference (Figure 7B, C) but were absent in comparisons between Tx2783 or RTx436 sequence and the BTx623 Bionano map (Figure 7D, E) and between Tx2783 sequence and RTx436 Bionano map (Figure 7F), suggesting potential artifacts in the BTx623 reference assembly. A total of 36 large inversions (>100 kb) were detected in the two lines in the same area (Supplementary Table S6). The pseudomolecules of the reference genome BTx623 were constructed by genetic maps (51). Several potential inversions detected in the comparison between the BTx623 Bionano map and the sorghum v3.1 reference sequence were located on chromosomes 5, 6 and 7, near the pericentromeric region as defined by genetic maps. While inversions within pericentromeric regions are recognized for their prevalence and significant evolutionary implications (73,74), our analysis suggests that these inversions may represent potential orientation errors in the reference genome. This observation was further supported and confirmed by the published RTx430 genome (75), which utilized whole genome shotgun (WGS) reads from RTx430 to validate inversion breakpoints.

BTx623 was used to characterize single-nucleotide variants (SNVs or SNPs) and INDELs in the Tx2783 and RTx436. These variations were classified based on their location as either intergenic or genic. In total, Tx2783 exhibited 1 223 983 intergenic SNPs and 233 153 genic SNPs, whereas RTx436 displayed a slightly higher number of intergenic SNPs at 1 556 765 and a lower count of genic SNPs at 225 254. Similarly, Tx2783 had 244 513 intergenic INDELs and 72 528

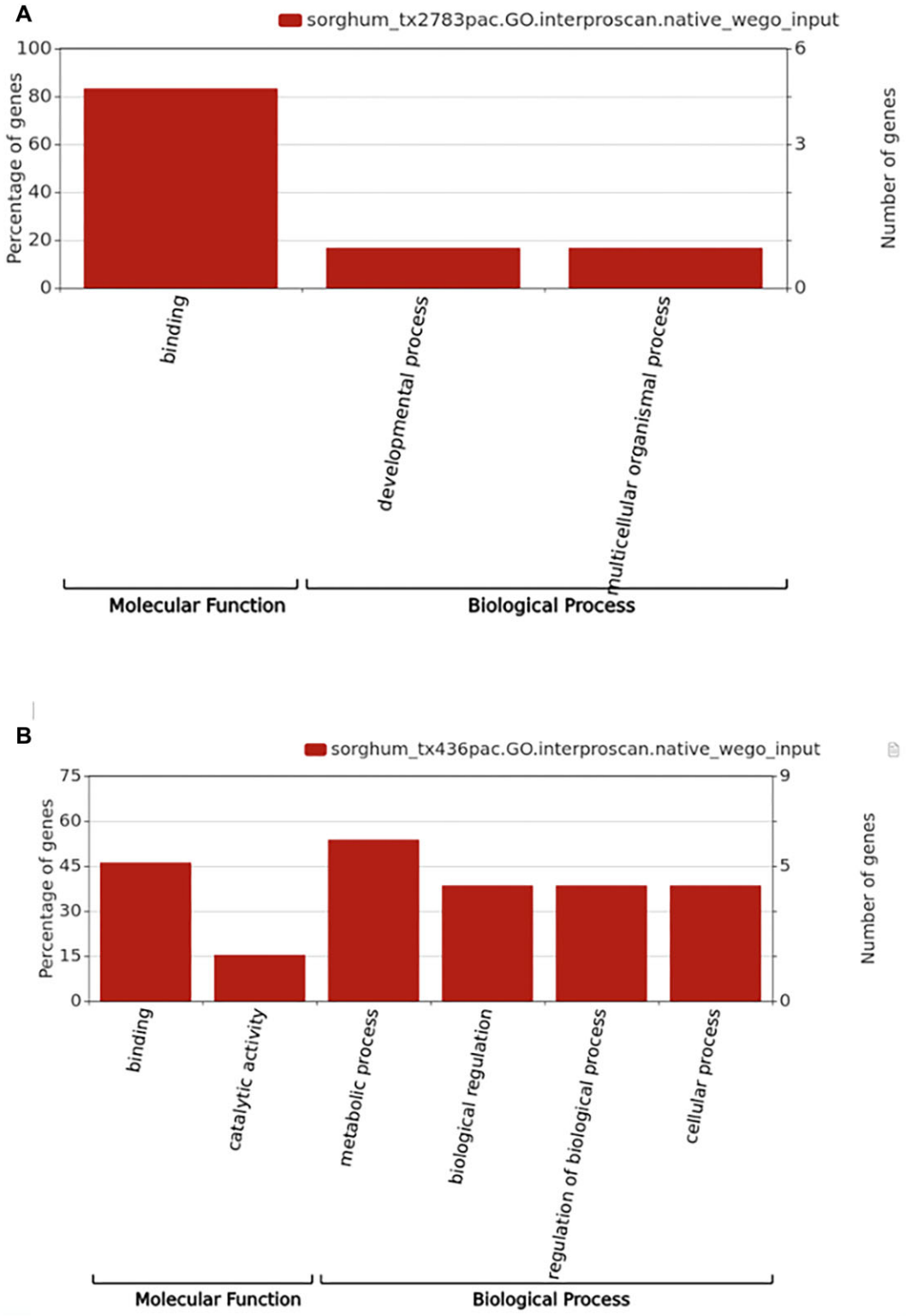

Figure 3. (**A**) Gene Ontology enrichment of genes that are unique to Tx2783. (**B**) Gene Ontology enrichment of genes that are unique to RTx436

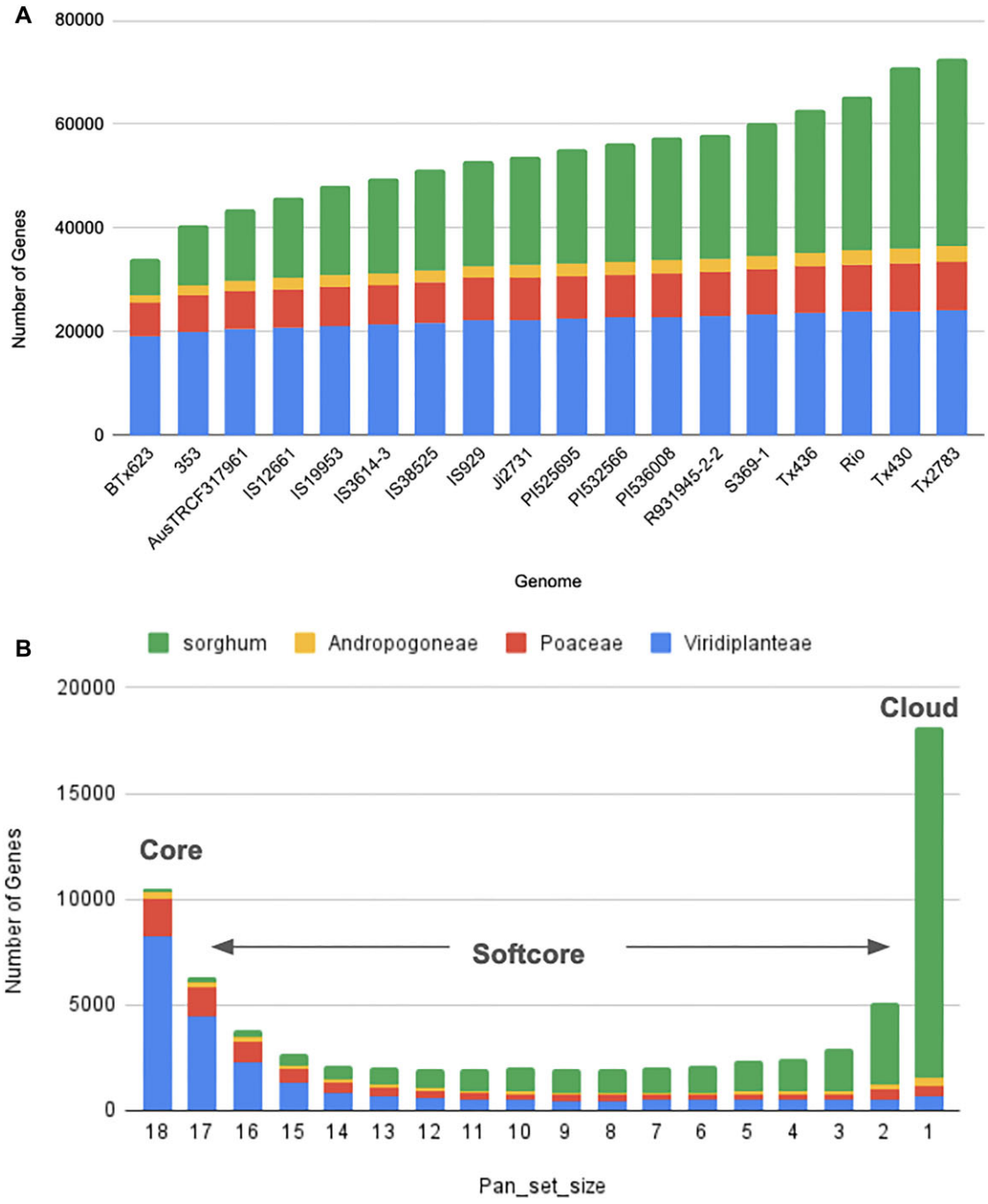

Figure 4. Pan-gene index using 18 sorghum genome annotations. (**A**) Pan gene set growth age of the gene. Core genes remain stable while increase in pan set due to lineage specific genes. (**B**) Distribution of core (18), softcore (2–17) and cloud (1) genes in the pan gene set. As expected core genes contain older conserved genes while softcore and cloud lineage specific or new evolving genes.

genic INDELs, while RTx436 showed a slightly higher number of intergenic INDELs at 266 120 and a lower count of genic INDELs at 67 427 (Supplementary Table S7). The number of SNPs and INDELs detected was similar to that reported in other studies involving sorghum accessions (59).

Discussion

Sorghum is a very important cereal grain, forage, and bioenergy crop around the world. Understanding the genetic diversity within sorghum provides a roadmap for improving this crop. The first reference genome was released in 2009 (3) and was updated 9 years later (51). With the development of long-read sequencing technology, many other sorghum genomes have become available to the public. Very recently, a sorghum pan-genome was constructed using assembled genomes representing cultivated and wild relatives (59).

In our study, we provide two high-quality sorghum genomes Tx2783 and RTx436 for the sorghum community with contig N50 19-fold and 11-fold higher compared to existing Btx623 v3.0 reference genome. In this study, Bionano Genomics DLS optical maps were employed. DLS opti-

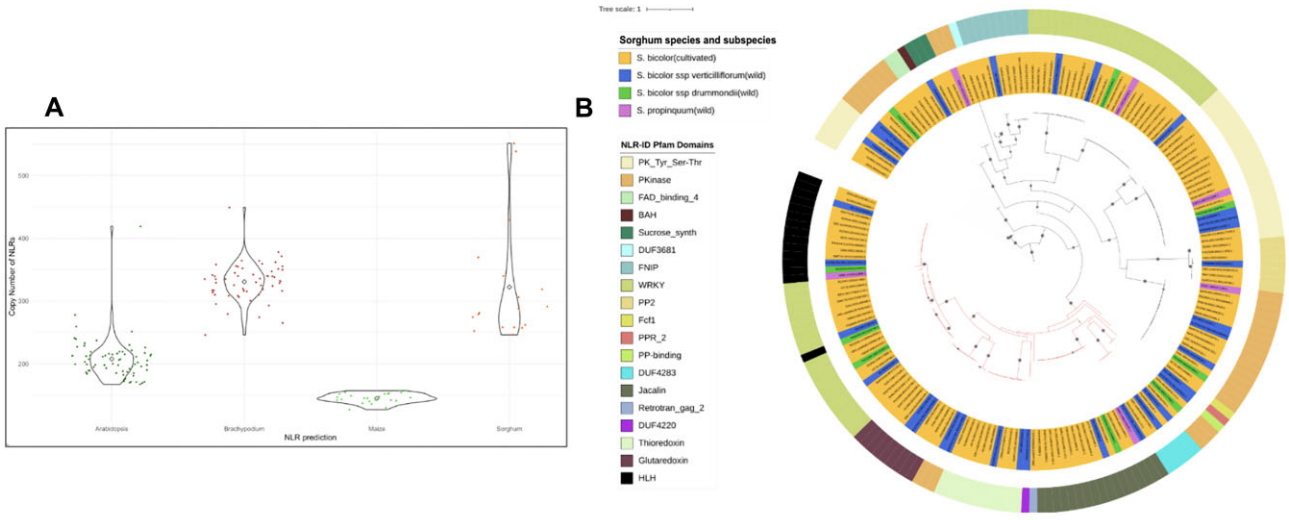

Figure 5. (**A**) Violin plot of NLR variation in pan-genomes of monocot species (*Z. mays, S. bicolor* and *B. distachyon*) and eudicot (*A. thaliana*). (**B**) Maximum likelihood phylogeny of NLRs containing integrated domains from sorghum lines. Dots indicate bootstrap values >80. Outer ring indicates the additional non-canonical domain present in the NLR-ID. Inner ring represents the type group the sorghum line belongs to.

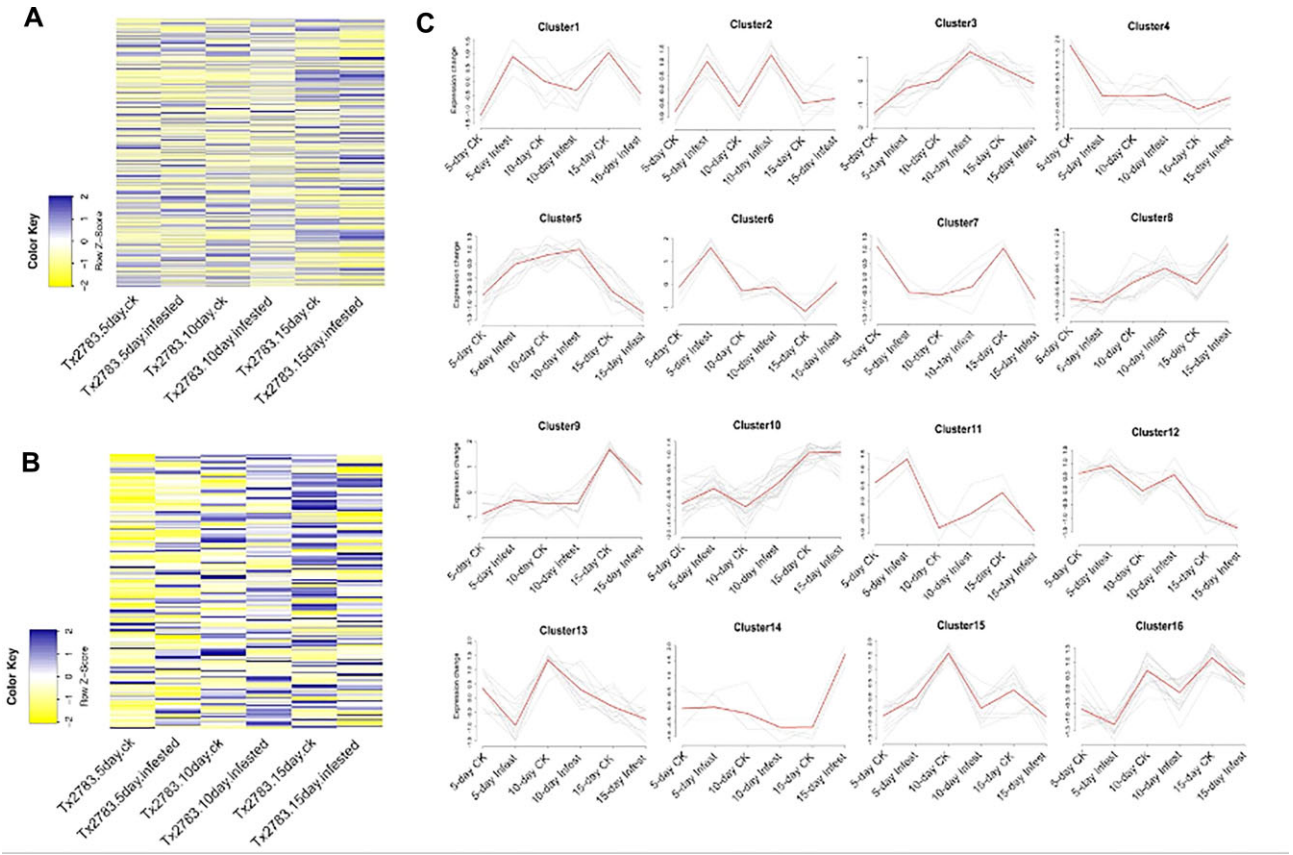

Figure 6. Gene expression analysis after sugarcane aphid infestation. (**A**) Heatmap of genes expressed before and after sugarcane aphid infestation. (**B**) Heatmap of NLR genes before and after sugarcane aphid infestation. (**C**) Clusters of genes expressed before and after sugarcane aphid infestation

cal maps are known for their considerable length, aiding in the correction, orientation, and hybrid scaffolding of PacBio polished contigs. During the hybrid scaffolding process, chimeric mis-assemblies within the contigs were detected and rectified to resolve conflicts with the genome maps. Alignment of the DLS optical maps generated in this study with the S. bicolor reference genome revealed potential large inversions in the Btx623 genome when compared to Tx2783 and RTx436, particularly within the pericentromeric region rich in repetitive DNA sequences crucial for cell division. However, additional analysis will be required for confirmation.

Despite significant improvements in genome quality facilitated by long-read sequencing technology, certain complex regions, including recently tandem-duplicated sequences, may

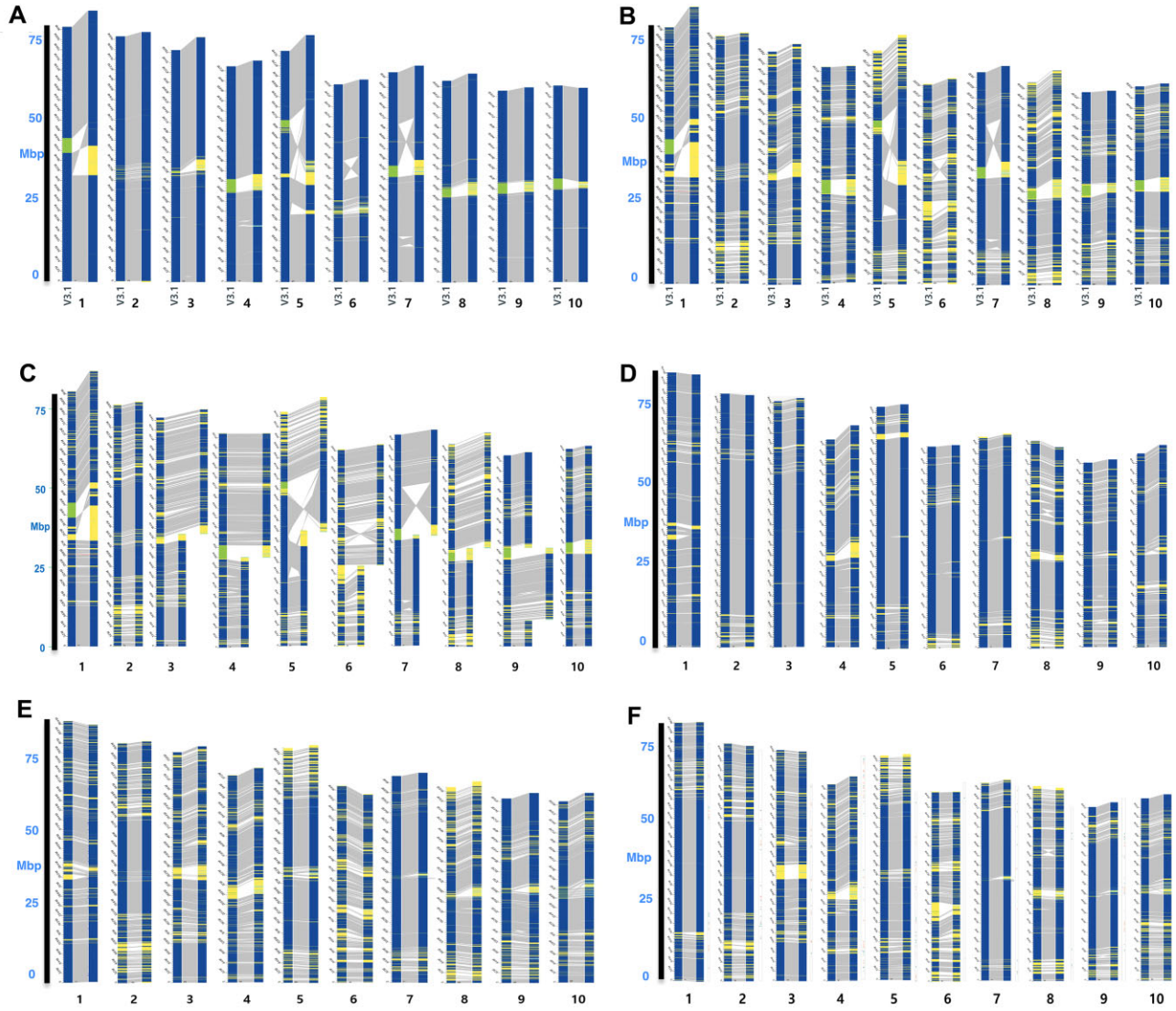

Figure 7. Alignments of Bionano maps between sorghum genomes. (**A**) BTx623 v3.1 reference sequence and BTx623 Bionano map; (**B**) BTx623 v3.1 reference sequence and RTx436 Bionano map; (**C**) BTx623 v3.1 reference sequence and Tx2783 Bionano map; (**D**) BTx623 Bionano map and Tx2783 sequence; (**E**) BTx623 Bionano map and RTx436 sequence; (**F**) Tx2783 sequence and RTx436 Bionano map.

require further refinement due to potential misassembly and collapse. This underscores the importance of integrating optical maps in constructing high-quality genome assemblies. The SNP and INDEL variations in the Tx2783 and RTx436 sorghum genomes were characterized using BTx623 as the reference. The number of SNPs and INDELs detected was similar to that reported in other studies involving sorghum accessions (59).

Our comparative analysis using canonical proteins from 26 plant genomes resulted in 31 210 protein-coding gene families of which ∼14K are sorghum specific gene trees. We were able to construct a pan gene index using 18 reference genome assemblies, and built a pangene index of 72 579 pan-genes sorghum genes and classified them as core, softcore or dispensable and found 75% of the pan-genes were segregating with the other accessions.

Characterization of disease resistance genes using genomic approaches offers an opportunity to study mechanisms of host pathogen interaction. Here we analyzed proteins for NLR genes in Sorghum annotations and found enrichment of NLR on chromosomes 5 and 8 with ∼6% of the NLRs showing presence of integrated domains. In sorghum, sugarcane aphid has become an aggressive pest of sorghum, causing severe yield losses. A high quality genome sequence of the sugarcane aphid resistance line will serve as an important resource for the sorghum research community to identify candidate genes and genomic regions associated with the sugarcane aphid resistance response.

Data availability

The PacBio and Bionano data of the sorghum Tx2783 and RTx436 genomes generated in this study have been deposited in the European Nucleotide Archive (ENA) under accession numbers ERS4546874 and ERS4546867, respectively. 10X Chromium sequencing data are available under accession ERS4804287 and ERS4804288 for Tx2783 and RTx436 respectively. Genome assemblies of Tx2783 and RTx436 are available under accession numbers GCA_903 166 285

and GCA_903 166 325, respectively. RAMPAGE datasets are available under accession number PRJEB42222.

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

We thank John Burke, Geoffrey Morris, Bob Klein, Bill Rooney, and Gary Peterson for helpful input concerning germplasm selection. The authors thank Peter Van Buren for help with the computational system. We acknowledge Peter VanBuren for systems support at Cold Spring Harbor Laboratory.

Author contributions: Bo Wang: Conceptualization, Formal analysis, Methodology, Validation, Writing—original draft.,Kapeel Chougule: Conceptualization, Formal analysis, Methodology, Validation, Writing—original draft.,Yinping Jiao: Conceptualization, Writing—review editing.,Andrew Olson: Formal analysis,Vivek Kumar: Formal analysis, Writing—review editing.,Nicholas Gladman: Formal analysis, Jian Huang: Formal analysis, Victor Llaca: Formal analysis, Kevin Fengler: Formal analysis, Xuehong Wei: Formal analysis, Liya Wang: Formal analysis, Xiaofei Wang: Formal analysis, Michael Regulski: Data Curation, Methodology, Jorg Drenkow: Data Curation, Methodology,Thomas Gingeras: Data Curation, Methodology, Chad Hayes: Data Curation, J. Scott Armstrong: Conceptualization, Yinghua Huang: Data Curation, Methodology, Zhanguo Xin: Conceptualization, Supervision, Writing—review editing, Doreen Ware: Conceptualization, Supervision, Writing—review editing.

Funding

United States Department of Agriculture-Agriculture Research Service (USDA-ARS) [8062-21000-051-000D; 8062-21000-044-000D; 3070-21000-009-00D]; National Science Foundation NSF Gramene [52930511]; Elzar High Performance Computing facility [NIH S10 OD0286321-01] at Cold Spring Harbor Laboratory.

Conflict of interest statement

No funding body had any role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript. None of the authors has any competing interests.

References

1. Ordonio,R., Ito,Y., Morinaka,Y., Sazuka,T. and Matsuoka,M. (2016) Molecular breeding of sorghum bicolor, a novel energy crop. *Int. Rev. Cell Mol. Biol.*, **321**, 221–257.
2. Morris,G.P., Ramu,P., Deshpande,S.P., Hash,C.T., Shah,T., Upadhyaya,H.D., Riera-Lizarazu,O., Brown,P.J., Acharya,C.B., Mitchell,S.E., *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.
3. Paterson,A.H., Bowers,J.E., Bruggmann,R., Dubchak,I., Grimwood,J., Gundlach,H., Haberer,G., Hellsten,U., Mitros,T., Poliakov,A., *et al.* (2009) The sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551–556.
4. Mace,E.S., Tai,S., Gilding,E.K., Li,Y., Prentis,P.J., Bian,L., Campbell,B.C., Hu,W., Innes,D.J., Han,X., *et al.* (2013) Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nat. Commun.*, **4**, 2320.
5. Mbulwe,L., Peterson,G.C., Scott-Armstrong,J. and Rooney,W.L. (2016) Registration of Sorghum germplasm Tx3408 and Tx3409 with tolerance to sugarcane aphid [Melanaphis sacchari (Zehntner)]. *Jo. Plant Registrations*, **10**, 51–56.
6. Luo,M. and Wing,R.A. (2003) An improved method for plant BAC library construction. *Methods Mol. Biol.*, **236**, 3–20.
7. Koren,S., Walenz,B.P., Berlin,K., Miller,J.R., Bergman,N.H. and Phillippy,A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive -mer weighting and repeat separation. *Genome Res.*, **27**, 722–736.
8. Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.
9. Li,J., Singh,U., Bhandary,P., Campbell,J., Arendsee,Z., Seetharam,A.S. and Wurtele,E.S. (2021) Foster thy young: enhanced prediction of orphan genes in assembled genomes. *Nucleic Acids Res.*, **50**, e37.
10. Grabherr,M.G., Haas,B.J., Yassour,M., Levin,J.Z., Thompson,D.A., Amit,I., Adiconis,X., Fan,L., Raychowdhury,R., Zeng,Q., *et al.* (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.*, **29**, 644–652.
11. Pertea,M., Pertea,G.M., Antonescu,C.M., Chang,T.-C., Mendell,J.T. and Salzberg,S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290–295.
12. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
13. Song,L., Sabunciyan,S., Yang,G. and Florea,L. (2019) A multi-sample approach increases the accuracy of transcript assembly. *Nat. Commun.*, **10**, 5000.
14. Venturini,L., Caim,S., Kaithakottil,G.G., Mapleson,D.L. and Swarbreck,D. (2018) Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *Gigascience*, **7**, giy093.
15. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
16. Wu,T.D. and Watanabe,C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.
17. Mapleson,D., Venturini,L., Kaithakottil,G. and Swarbreck,D. (2018) Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. *Gigascience*, **7**, giy131.
18. Altschul,S. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
19. Haas,B.J., Delcher,A.L., Mount,S.M., Wortman,J.R., Smith,R.K. Jr, Hannick,L.I., Maiti,R., Ronning,C.M., Rusch,D.B., Town,C.D., *et al.* (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.*, **31**, 5654–5666.
20. Wang,B., Regulski,M., Tseng,E., Olson,A., Goodwin,S., McCombie,W.R. and Ware,D. (2018) A comparative transcriptional landscape of maize and sorghum obtained by single-molecule sequencing. *Genome Res.*, **28**, 921–932.
21. Abdel-Ghany,S.E., Hamilton,M., Jacobi,J.L., Ngam,P., Devitt,N., Schilkey,F., Ben-Hur,A. and Reddy,A.S.N. (2016) A survey of the sorghum transcriptome using single-molecule long reads. *Nat. Commun.*, **7**, 11706.
22. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
23. Campbell,M.S., Law,M., Holt,C., Stein,J.C., Moghe,G.D., Hufnagel,D.E., Lei,J., Achawanantakun,R., Jiao,D., Lawrence,C.J.,

et al. (2014) MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.*, **164**, 513–524.
24. Holt,C. and Yandell,M. (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinf.*, **12**, 491.
25. Zhang,R.-G., Li,G.-Y., Wang,X.-L., Dainat,J., Wang,Z.-X., Ou,S. and Ma,Y. (2022) TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hortic Res.*, **9**, uhac017.
26. Neumann,P., Novák,P., Hoštáková,N. and Macas,J. (2019) Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mob. DNA*, **10**, 1.
27. Hoff,K.J., Lange,S., Lomsadze,A., Borodovsky,M. and Stanke,M. (2016) BRAKER1: unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, **32**, 767–769.
28. Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.
29. Hufford,M.B., Seetharam,A.S., Woodhouse,M.R., Chougule,K.M., Ou,S., Liu,J., Ricci,W.A., Guo,T., Olson,A., Qiu,Y., *et al.* (2021) De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*, **373**, 655–662.
30. International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
31. International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature*, **463**, 763–768.
32. Cheng,C.Y., Krishnakumar,V., Chan,A.P., Thibaud-Nissen,F., Schobel,S. and Town,C.D. (2017) Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. *Plant J.*, **89**, 789–804.
33. Jones,P., Binns,D., Chang,H.-Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G., *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
34. Olson,A.J. and Ware,D. (2021) Ranked choice voting for representative transcripts with TRaCE. *Bioinformatics*, **38**, 261–264.
35. Stabenau,A., McVicker,G., Melsopp,C., Proctor,G., Clamp,M. and Birney,E. (2004) The Ensembl core software libraries. *Genome Res.*, **14**, 929–933.
36. Batut,P. and Gingeras,T.R. (2013) RAMPAGE: promoter activity profiling by paired-end sequencing of 5'-complete cDNAs. *Curr. Protoc. Mol. Biol.*, **104**, Unit 25B.11.
37. Wang,L., Lu,Z., Van Buren,P. and Ware,D. (2018) SciApps: a cloud-based platform for reproducible bioinformatics workflows. *Bioinformatics*, **34**, 3917–3920.
38. Frith,M.C., Valen,E., Krogh,A., Hayashizaki,Y., Carninci,P. and Sandelin,A. (2008) A code for transcription initiation in mammalian genomes. *Genome Res.*, **18**, 1–12.
39. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
40. Ramírez,F., Ryan,D.P., Grüning,B., Bhardwaj,V., Kilpert,F., Richter,A.S., Heyne,S., Dündar,F. and Manke,T. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.*, **44**, W160–W165.
41. Gladman,N., Olson,A., Wei,S., Chougule,K., Lu,Z., Tello-Ruiz,M., Meijs,I., Van Buren,P., Jiao,Y., Wang,B., *et al.* (2022) SorghumBase: a web-based portal for sorghum genetic information and community advancement. *Planta*, **255**, 35.
42. Vilella,A.J., Severin,J., Ureta-Vidal,A., Heng,L., Durbin,R. and Birney,E. (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.
43. Jones,J.D.G. and Dangl,J.L. (2006) The plant immune system. *Nature*, **444**, 323–329.
44. Jacob,F., Vernaldi,S. and Maekawa,T. (2013) Evolution and conservation of plant NLR functions. *Front. Immunol.*, **4**, 297.
45. Sarris,P.F., Cevik,V., Dagdas,G., Jones,J.D.G. and Krasileva,K.V. (2016) Comparative analysis of plant immune receptor architectures uncovers host proteins likely targeted by pathogens. *BMC Biol.*, **14**, 8.
46. van der Biezen,E.A. and Jones,J.D. (1998) The NB-ARC domain: a novel signalling motif shared by plant resistance gene products and regulators of cell death in animals. *Curr. Biol.*, **8**, R226–R227
47. Wickham,H. (2016) In: *Ggplot2: Elegant Graphics for Data Analysis*. Springer.
48. Letunic,I. and Bork,P. (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.*, **44**, W242–W245.
49. Tetreault,H.M., Grover,S., Scully,E.D., Gries,T., Palmer,N.A., Sarath,G., Louis,J. and Sattler,S.E. (2019) Global responses of resistant and susceptible sorghum (Sorghum bicolor) to sugarcane aphid (Melanaphis sacchari). *Front. Plant Sci.*, **10**, 145.
50. Kumar,L. and E Futschik,M. (2007) Mfuzz: a software package for soft clustering of microarray data. *Bioinformation*, **2**, 5–7.
51. McCormick,R.F., Truong,S.K., Sreedasyam,A., Jenkins,J., Shu,S., Sims,D., Kennedy,M., Amirebrahimi,M., Weers,B.D., McKinley,B., *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.
52. Ou,S., Chen,J. and Jiang,N. (2018) Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.*, **46**, e126.
53. Simão,F.A., Waterhouse,R.M., Ioannidis,P., Kriventseva,E.V. and Zdobnov,E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.
54. Ou,S., Su,W., Liao,Y., Chougule,K., Agda,J.R.A., Hellinga,A.J., Lugo,C.S.B., Elliott,T.A., Ware,D., Peterson,T., *et al.* (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.*, **20**, 275.
55. Tello-Ruiz,M.K., Naithani,S., Gupta,P., Olson,A., Wei,S., Preece,J., Jiao,Y., Wang,B., Chougule,K., Garg,P., *et al.* (2021) Gramene 2021: harnessing the power of comparative genomics and pathways for plant research. *Nucleic Acids Res.*, **49**, D1452–D1463.
56. Smith,A.B. and Jones,C.D. (2018) The role of transposable elements in plant development. *Plant Biol.*, **20**, 123–135.
57. Niu,D., Lin,X.-L., Kong,X., Qu,G.-P., Cai,B., Lee,J. and Jin,J.B. (2019) SIZ1-Mediated SUMOylation of TPR1 suppresses plant immunity in arabidopsis. *Mol. Plant*, **12**, 215–228.
58. Shang,L., Li,X., He,H., Yuan,Q., Song,Y., Wei,Z., Lin,H., Hu,M., Zhao,F., Zhang,C., *et al.* (2022) A super pan-genomic landscape of rice. *Cell Res.*, **32**, 878–896.
59. Tao,Y., Luo,H., Xu,J., Cruickshank,A., Zhao,X., Teng,F., Hathorn,A., Wu,X., Liu,Y., Shatte,T., *et al.* (2021) Extensive variation within the pan-genome of cultivated and wild sorghum. *Nat. Plants*, **7**, 766–773.
60. Gordon,S.P., Contreras-Moreira,B., Woods,D.P., Des Marais,D.L., Burgess,D., Shu,S., Stritt,C., Roulin,A.C., Schackwitz,W., Tyler,L., *et al.* (2017) Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nat. Commun.*, **8**, 2184.
61. Stein,J.C., Yu,Y., Copetti,D., Zwickl,D.J., Zhang,L., Zhang,C., Chougule,K., *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.
62. Zhou,Y., Chebotarov,D., Kudrna,D., Llaca,V., Lee,S., Rajasekar,S., Mohammed,N., *et al.* (2020) A platinum standard pan-genome resource that represents the population structure of Asian rice. *Sci. Data*, **7**, 113.

63. Jain,R., Jenkins,J., Shu,S., Chern,M., Martin,J.A., Copetti,D., *et al.* (2019) Genome sequence of the model rice variety KitaakeX. *Bmc Genomics [Electronic Resource]*, **20**, 905.
64. Song,J.-M., *et al.* (2021) Two gap-free reference genomes and a global view of the centromere architecture in rice. *Mol. Plant*, **14**, 1757–1767.
65. Swigonová,Z., Lai,J., Ma,J., Ramakrishna,W., Llaca,V., Bennetzen,J.L. and Messing,J. (2004) Close split of sorghum and maize genome progenitors. *Genome Res.*, **14**, 1916–1923.
66. Hassani-Pak,K., Knet,M. and Grüning,B. (2020) Heterogeneous annotation workflows for plant genomes. *Trends Plant Sci.*, **25**, 694–707.
67. Cook,D.E., Mesarich,C.H. and Thomma,B.P. (2015) Understanding plant immunity as a surveillance system to detect invasion. *Annu. Rev. Phytopathol.*, **53**, 541–563.
68. Jones,J.D.G., Vance,R.E. and Dangl,J.L. (2016) Intracellular innate immune surveillance devices in plants and animals. *Science*, **354**, aaf6395.
69. Kourelis,J. and van der Hoorn,R.A. (2018) Defended to the nines: 25 years of resistance gene cloning identifies nine mechanisms for R protein function. *Plant Cell*, **30**, 285–299.
70. Dangl,J.L. and Jones,J.D.G. (2001) Plant pathogens and integrated defence responses to infection. *Nature*, **411**, 826–833.
71. Han,Q., Zhang,J., Li,H., Luo,Z., Ziaf,K., Ouyang,S. and Zhang,Z. (2012) Identification and expression pattern of one stress-responsive NAC gene from Solanum lycopersicum. *Mol. Biol. Rep.*, **39**, 6285–6292.
72. Bailey,P.C., Schudoma,C., Jackson,W., Baggs,E., Dagdas,G., Haerty,W., Moscou,M. and Krasileva,K.V. (2018) Dominant integration locus drives continuous diversification of plant immune receptors with exogenous domain fusions. *Genome Biol.*, **19**, 23.
73. Cheng,Y., Ma,Y. and Li,J. (2018) The pericentromeric region: a focus for chromosome evolution. *Plant J.*, **95**, 659–672.
74. Melters,D.P., Bradnam,K.R., Young,H.A., Telis,N., May,M.R., Ruby,J.G. and Smith,A.D. (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol.*, **14**, R10.
75. Deschamps,S., Zhang,Y., Llaca,V., Ye,L., Sanyal,A., King,M., May,G. and Lin,H. (2018) A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nat. Commun.*, **9**, 4844.

Received: September 12, 2023. **Revised:** July 1, 2024. **Editorial Decision:** July 17, 2024. **Accepted:** July 23, 2024
Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics 2024.