bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Pan-genome Analysis in Sorghum Highlights the Extent of Genomic Variation and Sugarcane Aphid Resistance Genes

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

[1] Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA

[2] Texas Tech University, 1006 Canton Ave, Lubbock, TX 79409-2122

[3] Department of Plant and Soil Sciences, Oklahoma State University, Stillwater, OK, 74078-6028, USA

[4] Corteva Agriscience™, 8325 NW 62nd Avenue, Johnston, IA, 50131, USA

[5] U.S. Department of Agriculture-Agricultural Research Service, Plant Stress and Germplasm Development Unit, Cropping Systems Research Laboratory, Lubbock, TX, 79415, USA

[6] Wheat, Peanut and Other Field Crops Research Unit, 1301 N. Western Rd. Stillwater, OK, 74075, USA

[7] USDA-ARS Plant Science Research Laboratory, 1301 N. Western Road, Stillwater, OK 74075-2714, USA

[8] Dept. of Plant Biology, Ecology, and Evolution, 301 Physical Sciences, Stillwater, OK 74078-3013, USA

[9] USDA ARS Robert W. Holley Center for Agriculture and Health Cornell University, Ithaca, New York, USA

*Author for correspondence:

Doreen Ware: ware@cshl.edu

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ABSTRACT

Sorghum bicolor, one of the most important grass crops around the world, harbors a high degree of genetic diversity. We constructed chromosome-level genome assemblies for two important sorghum inbred lines, Tx2783 and RTx436. The final high-quality reference assemblies consist of 19 and 18 scaffolds, respectively, with contig N50 values of 25.6 and 20.3 Mb. Genes were annotated using evidence-based and *de novo* gene predictors, and RAMPAGE data demonstrate that transcription start sites were effectively captured. Together with other public sorghum genomes, BTx623, RTx430, and Rio, extensive structural variations (SVs) of various sizes were characterized using Tx2783 as a reference. Genome-wide scanning for disease resistance (R) genes revealed high levels of diversity among these five sorghum accessions. To characterize sugarcane aphid (SCA) resistance in Tx2783, we mapped the resistance region on chromosome 6 using a recombinant inbred line (RIL) population and found a SV of 191 kb containing a cluster of R genes in Tx2783. Using Tx2783 as a backbone, along with the SVs, we constructed a pan-genome to support alignment of resequencing data from 62 sorghum accessions, and then identified core and dispensable genes using this population. This study provides the first overview of the extent of genomic structural variations and R genes in the sorghum population, and reveals potential targets for breeding of SCA resistance.

Key words: pan-genome, structural variation, sugarcane aphid, resistance

INTRODUCTION

Sorghum bicolor, the fifth most economically important cereal, forage, and bioenergy crop in the world (Ordonio et al., 2016), 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 (Morris et al., 2013). Moreover, its small and compact genome relative to other C4 grasses makes it an excellent model for genomic studies (Paterson et al., 2009). Previous

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 (Morris et al., 2013). 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 (Mace et al., 2013).

Previous studies of sorghum diversity were mainly limited to single-nucleotide polymorphisms (SNPs) and small insertions/deletions (indels) called from short-read sequencing technologies (Mace et al., 2013; Brenton et al., 2016; Zarrei et al., 2015; Saxena et al., 2014). Recent development of long-read sequencing technologies have brought genomics studies into a new era for many species, including maize (Jiao et al., 2017), rice (Zhang et al., 2016), *Arabidopsis* (Michael et al., 2018), and sorghum (Deschamps et al., 2018). However, the diversity within a species cannot be represented by a single reference genome, and this realization has led to the spread of the pan-genome concept. A pan-genome represents the genomic diversity of a species and includes core genes found in all individuals, as well as variable genes that are absent from some individuals (Bayer et al., 2020). In addition, genomic structural variations (SVs), including deletions, insertions, inversions, and duplications, represent an important source of genetic diversity (Alkan et al., 2011; Baker, 2012; Zhang et al., 2015), and such variations have been implicated in several human diseases (Sebat et al., 2007; Stefansson et al., 2008). In plants, SVs are associated with variations in phenotypes including leaf size (Horiguchi et al., 2009), fruit shape (Xiao et al, 2008), and aluminum tolerance (Maron et al., 2013). Accurate identification and characterization of large SVs will require long reads and high-quality genome assemblies. Long reads are advantageous for SV calling because they can span repetitive or those that are problematic for other reasons (Mahmoud et al., 2019). One class of gene-space structural variants are the disease-related nucleotide-binding site plus leucine-rich repeat (NBS-LRR) genes, which in sorghum have higher diversity than non–NBS-LRR genes (Mace et al., 2014). Several NBS-LRR genes have also been identified in *Arabidopsis* (Meyers et al., 2003) and rice (Stein et al., 2018; Zhou et al., 2004).

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 (SCA) resistance, and RTx436, another widely utilized pollinator parent with known general combinability. We used the new reference genomes and three publicly available accessions, BTx623 (McCormick et al., 2018), RTx430 (Deschamps et al., 2018), and Rio (Cooper et al., 2019), to characterize SVs. Using a recombinant inbred line (RIL) population between Tx2783 and BTx623, we identified an SV on chromosome 6 that segregates with a QTL identified for SCA tolerance. In addition, we constructed a sorghum pan-genome using TX2783 as the backbone, along with insertions called from the four complete genomes, and used it to identify core and dispensable genes based on resequencing data from 62 sorghum accessions. Our results provide new insights into sorghum population diversity, as well as potential targets for breeding sorghum for high sugarcane aphid tolerance.

RESULTS

Chromosome-level Genome Assembly of Two Sorghum Inbred Lines

To characterize large SVs in the sorghum population and provide support for modern breeding, we selected two important sorghum inbred lines, Tx2783 and RTx436, for genome construction. Line Tx2783 (PI 656001) was originally bred for resistance to sorghum greenbug (*Schizaphis graminum*) and also exhibits high SCA resistance (Tetreault et al., 2019). 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 20.3 Mb for RTx436. In addition, we generated Bionano molecules for Tx2783 and RTx436 that

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

yielded genome maps of 721.504 Mb and 723.680 Mb, respectively, with N50 lengths of 36.987 Mb and 37.781 Mb (Supplemental Table S1). The chromosomes of these two genomes were constructed using hybrid scaffolds generated from Bionano genome maps. Most of the chromosomes consisted of two scaffolds (Fig. 1A, Table 1). These two genomes were of higher quality than the published sorghum reference genomes of BTx623 (McCormick et al., 2018) and RTx430 (Deschamps et al., 2018), with 15-fold and 7-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. Gene space assessment using BUSCO confirmed the completeness of the Tx2783 and RTx436 genome assemblies (Fig.1B).

Transposable Element and Gene Annotation of the Tx2783 and RTx436 Genomes

Transposon elements (TEs) were annotated using EDTA (Ou et al., 2019), with 68.97%, 69.34%, and 69.35% of genome sequences 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 36.63%, 37.20%, and 37.98% of TEs in the Tx2783, RTx436, and BTx623 genomes, respectively (Supplemental Table S2). In addition, we also identified 10984 (79.5%), 11207 (79.7%), and 10746 (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 Methods), and then filtered based on annotation evidence distance (AED) score (Campbell et al., 2014) and homology to a maize, *Brachypodium*, rice, or *Arabidopsis* protein sequences obtained from Gramene release 62 (Tello-Ruiz et al., 2020). Ultimately, this approach generated a total of 29,612 and 29,265 protein-coding genes and 4,205 and 3,478 non-coding genes in the Tx2783 and RTx436 genomes, respectively (Supplemental Table S3). However, the RTx436 annotation predicted slightly longer genes with a higher intron count; accordingly, Tx2783 had slightly more single-exon genes than RTx436. 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,

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

root and shoot tissues from Tx2783 and RTx436 were collected for RAMPAGE (RNA annotation and mapping of promoters for analysis of gene expression) assays (Batut et al., 2013). This analysis identified 228,249 and 300,633 high-confidence peaks from Tx2783 and RTx436 respectively. The narrow distribution of RAMPAGE signals (Fig. 2A and 2B) 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.

Characterization of Genetic Variations in Five Sorghum Reference Assemblies

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 (Fig. 3A), which were also found in the alignment between Tx2783, RTx436, and the v3.1 reference (Fig. 3B and 3C) but were absent in comparisons between Tx2783 or RTx436 and the BTx623 Bionano map (Fig. 3D and 3E) and between Tx2783 and RTx436 (Fig. 3F), suggesting potential artifacts in the public BTx623 reference assembly. A list of these SVs from different alignments is provided in Supplemental Data S1. A total of 36 large inversions (>100kb) were detected in the two lines in the same area (Supplemental Table S4). The pseudomolecules of the reference genome BTx623 was constructed by genetic maps (McCormick et al., 2018). As most of the large inversions were in the regions with low recombination rate in the genetic maps, it suggested that these inversions might be the orientation errors in the BTx623 reference genome. Similar observation was also reported from the published RTx430 genome (Deschamps et al., 2018).

Based on this assessment, for the remaining analyses Tx2783 was used to characterize single-nucleotide variants (SNVs or SNPs) in the BTx623, Rio, RTx436, and RTx430 genomes (Fig. 4A). RTx430 had the most SNPs among the four genomes (Supplemental Table S5). The bulk of the SNPs, 92.6% on average, were predominantly called in intergenic regions; however, in gene regions, SNPs within the CDS regions of exons, which were the most highly conserved, were likely to have large effects (Fig. 4B). To

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

identify large-effect SNPs within genes, we generated SIFT (sort intolerant from tolerant) scores using the Tx2783 genome annotation. The results revealed that 766, 1000, 1093, and 1004 genes in BTx623, Rio, RTx430, and RTx436, respectively, contained SNPs with SIFT scores < 0.05 and were therefore predicted to be deleterious mutations. Among these genes, 183 genes were shared by all four genomes (Fig. 4C). Gene Ontology analysis revealed that these genes were associated with molecular function terms related to transferase activity, catalytic activity, and binding activity and biological process terms related to nitrogen compound metabolic process (Fig. 4D, Pearson Chi-Square test, $P < 0.05$).

Although less common than SNPs, SVs have a greater potential to impact function due to their larger size and the possibility of altering gene structure, dosage, or location (Fuentes et al. 2019). We characterized SVs in BTx623, RTx436, RTx430, and Rio by whole-genome alignment using sorghum Tx2783 as the reference. Variants were classified into four different categories: insertion (INS), deletion (DEL), inversion (INV), and copy number variation (CNV). The largest number of SVs was identified in RTx430 (Figure 4A, Supplemental Table S6, Supplemental Data S2). Like the SNPs, these SVs were enriched in chromosome-terminal regions in all five genomes, but the specific SV categories that were enriched differed among individual genomes (Fig. 5A). In addition, we identified genotype-specific and common SVs in each genome (Fig. 5B to 5E), the majority of which were in intergenic regions; however, as noted above, SVs within genes were more likely to contribute to large-effect mutations (Fig. 5F–G). In addition, we also saw that the number of genes within SVs were correlated with the size of the SVs (Supplemental Fig. S1), and that overall, DEL SVs covered the most genes.

Combining the information on SNPs and SVs, we found that the number of SNPs was positively correlated with the number of INS and DEL SVs in the same genomic region; this correlation was lower for CNVs and absent for INV SVs (Supplemental Fig. S2). CNVs are a specific class of SVs that play important roles in plant development (Soyk et al., 2019). We also investigated CNVs in these genomes, and found that the number and size of CNVs were distributed similarly in each genome, although they were slightly

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

smaller and less abundant in the BTx623 genome (Supplemental Fig. S3A), and smaller CNVs were usually present in more copies than larger CNVs (Supplemental Fig. S3B, Supplemental Fig. S4).

We also investigated the possible functional roles of the genes affected by different types of SVs. Gene Ontology analysis using WEGO (Ye et al., 2018) revealed that genes affected by INS/DEL/CNVs were enriched for biological process terms related to metabolic and cellular processes. Enriched molecular functions included catalytic activity and binding (Supplemental Fig. S5).

Core and Dispensable Genes Identification Using 62 Genomes Resequencing Data

To assess the core and dispensable genes, we constructed a pan-genome using the Tx2783 assembly and novel genomic regions represented by the SV insertions from the four other assemblies (Supplemental Data S2). The pan-genome had an effective genome length of ~820 Mb and contained 35355 genes (29,612 protein coding + 4204 non-coding from TX2783 + 1539 [1517 protein coding and 22 non-coding] SVs annotated genes from Rio, RTx430, RTx436, and BTx623). Using publicly available resequencing data for 62 sorghum accessions (Supplemental Data S3; Mace et al., 2013; LeBauer et al., 2020; https://datacommons.cyverse.org/browse/iplant/home/shared/terraref), we aligned sequence reads were aligned to the sorghum pan-genome using BWA software (Li et al., 2009a). Mapping rates varied from 74.57% to 98.59%, averaging 93.25% in the resequencing lines (Supplemental Data S4), and were lowest in wild weedy lines (AusTRCF 317961, Greenleaf, PI330272, PI300119, Kilo and Zengada). The observed differences in mapping rates may be due to divergence between the sequenced genotypes. Using a conservative quality filter pipeline, we identified 8,052,666 SNPs throughout the genome in *S. bicolor* alone. SNPs were used to calculate the genetic distance between individuals, and a neighbour-joining tree was constructed using RAxML-NG (Kozlov et al., 2019) (Fig. 6A). In general, our phylogenetic tree was in consistent with a previous study by Mace et al.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Core and dispensable genes were identified by using genes covered in each of the 62 accessions under a 75% coverage cutoff. As a result, among the protein-coding genes, we identified 17,314 as core (present in all 62 lines), 3,466 as softcore (present in 59–61 lines), 9,952 as dispensable (present in 2–58 lines) and 173 as private genes (present in only one line) (Fig. 6B–C; Supplemental Table S7). Among the non-coding genes, we identified 1,803 as core, 734 as softcore, 1,906 as dispensable and 7 as private genes. (Supplemental Fig. S6, Supplemental Table S7).

NBS-LRR Genes in Sorghum Genomes

To characterize disease resistance (R) genes, we first assessed the distribution of known NBS-LRR genes in both maize and sorghum (Supplemental Fig. S7). The results revealed that maize has far fewer R genes than sorghum (Paterson et al., 2009; Schnable et al., 2009). Whole-genome scanning of each sorghum line revealed enrichment of R genes on chromosomes 5 and 8 (Fig. 7A), with RTx430 having the most R genes (363) and Rio having the fewest (319) (Supplemental Table S8).

In 2013, sugarcane aphid (SCA) emerged as a major insect pest of sorghum crops in North America (Alkan et al. 2011). The line Tx2783 is highly resistant to SCA, and thus has great potential to boost breeding targeted at aphid resistance (Tetreault et al., 2019) (Fig. 7B). Our genetic tests showed that the resistance allele in Tx2783 is dominant. To identify the resistance loci in Tx2783, we constructed a RIL population between Tx2783 and BTx623 and used it to generate two pools: one with 20 SCA-resistant individuals, and another with 20 susceptible individuals. After comparing the genotypes of the parental lines and the two RIL pools, we identified one peak on chromosome 6 harboring 17 large-effect variations as the candidate causal region in Tx2783 (Chr6: 3,008,428–3,286,874) (Fig. 7C, Supplemental Data S5). Our SV analysis revealed 11 deletions in this region in the susceptible line BTx623; the largest deletion was 191,539 bp, and the remaining deletions were relatively small (< 3000 bp). (Supplemental Table S9, Supplemental Fig. S8). We then checked the overlapping genes in this region in the Tx2783 genome, and found that eight genes overlapped with three deletions in this region (Supplemental Table S9). Functional annotation of these genes showed that six of them

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

encode proteins containing leucine-rich repeat (LRR) domains (Supplemental Data S6), which are involved in a variety of biological processes, including disease resistance and immune responses.

Genome sequence analysis in the sugarcane aphid resistance QTL region showed that there is a nearly identical 70,267-bp partial tandem duplication, which increased the complexity of this region (Supplemental Fig. S9) and required refinement of the loci (see Methods). Previous study has revealed that resistance genes are usually clustered within genomes (Stein et al., 2018). Therefore, we explored the depth of coverage of genes in the SV region in our population data (Fig. 7D). Based on the resequencing data, there was little to no coverage of this region in all but ten accessions for the eight genes. There were ten R genes in this region in Tx2783, but only three in BTx623. Therefore, we speculated that the gene dosage of these gene clusters in Tx2783 play a large role in SCA resistance.

To investigate how gene expression is impacted by SCA, we remapped the RNA-Seq data before and after sugarcane aphid infestation from a previous study (Tetreault et al., 2019) to our Tx2783 genome. Gene expression analysis revealed that five out of the ten R genes expressed before and after SCA infestation (Fig. 7E). We also observed diverse expression patterns for all genes throughout the genome (Fig. 8A). We then grouped all genes into 16 clusters based on their expression levels (Fig. 8B). This analysis revealed that 3,944 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 SCA 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 (Fig. 8C).

DISCUSSION

Sorghum is a very important cereal grain, forage, and bioenergy crop around the world. The first reference genome was released in 2009 (Paterson et al., 2009) and was updated 9 years later (McCormick et al., 2018). With the development of long-read sequencing

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

technology, two other sorghum genomes have become available to the public: RTx430 (Deschamps et al., 2018) and Rio (Cooper et al., 2019). To date, however, no studies have examined SVs in sorghum genomes at the pan-genome level.

In this study, we generated two long-read sequenced genomes: Tx2783, which is highly SCA-resistant, and RTx436, which is a high-yielding, well-adapted pollinator line. Using these and other available genomes, we were able to investigate SVs in sorghum. Using Bionano maps, we found several incorrectly assembled regions in the current BTx623 reference genome; these assembly errors were confirmed by various map alignments. In light of this, we used the higher-quality Tx2783 genome as reference for calling SVs for subsequent analyses. The results revealed a large number of SVs in various sorghum genomes, and also they share some SVs at the same locus using the same reference; moreover, we found that different genomes can have SVs at different regions.

As the first examination of SVs across several genomes in sorghum, this study will provide valuable information to sorghum pan-genome comparisons and deeper insights into plant genomics. It is worth noting that even though genome quality has been dramatically improved by long-read sequencing technology, some complex regions, including recently tandem duplicated sequences, may still need to be refined because these regions may incorrectly assembled and collapsed, as the SCA region was in this study. Visual inspection of the hybrid scaffold at the SCA locus revealed an approximately 69.8-kb collapsed repeat in the contig assembly that was not resolved during the initial manual curation process. Similarly, alignment of the PacBio corrected reads to the contig assembly showed evidence of a collapsed assembly of approximately 77.7 kb within the QTL locus. To address this issue, we manually assembled the longest corrected reads from a 592-kb target region spanning the QTL region using GeneCodes Sequencher, yielding a new sequence that resolved the collapsed repeat such that it was congruent to the Bionano map and incorporated the previously soft-clipped reads.

SCA has become an aggressive pest of sorghum, causing severe yield losses. Therefore, identification of SCA resistance loci is urgently required and has great potential value.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Tx2783 is a highly SCA-resistant line grown worldwide, and is thus an excellent resource for identification of resistance loci. In this study, we used a RIL population to map SCA resistance to a region on chromosome 6. We found several large insertions in this region in Tx2783 relative to the susceptible line BTx623. Eight genes overlapped with the SV region, of which six were in a 191-kb region, so it is reasonable to speculate that these genes or these SVs play important roles in SCA resistance. Our population analysis also showed that ten genes segregate with different sorghum accessions. We speculate that these ten gene clusters could function together in response to SCA as part of a complex resistance mechanism. Further studies in which one or all of these genes are mutated (e.g., using CRISPR technology) will be necessary to determine which genes confer the SCA-resistance phenotype, and to explore the underlying molecular mechanism. To broadly characterize R genes, we performed a whole-genome scan of NLR genes for each line. The results revealed that each sorghum genome has a similar number of R genes. Analysis of whole-genome distributions revealed that NLR genes are enriched on chromosomes 5 and 8; these R gene clusters could play important roles in resistance to SCA and other biotic stressors and diseases in sorghum. Our characterization of sorghum pan-genome variations and SCA resistance provides an important resource for the sorghum research community.

As mentioned above, genomics has entered the pan-genome era, in which the genetic diversity within a species is revealed through the availability of multiple genomes. On the other hand, the phenotypic differences among accessions are usually associated with dispensable and private genes. Consequently, it is important to distinguish the core and dispensable elements in a pan-genome. To construct a sorghum pan-genome, we used the Tx2783 genome as backbone and added insertions called from four other complete genomes. We then used genome resequencing data from 62 other sorghum accessions, including 44 accessions from Mace et al., 2013 and 17 additional BAP/SAP lines that add more diversity to the population, to identify the core and dispensable genes. After mapping these resequencing data to the pan-genome, SNPs were called and used to construct a phylogenetic tree, which was consistent with the findings of a previous study (Mace et al., 2013). We also observed that 21 lines had similar gene coverage among the

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

population, resulting in a blip in the cumulative curve of covered genes. We then identified core and dispensable genes from the population. These different classes of genes are likely to explain the genetic and phenotypic diversification of sorghum species, providing fundamental information for population genomics in sorghum.

METHODS

Data Generation

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 2 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 through 100- and 40-µm cell strainers, and then pelleted by centrifugation at 2,500 × *g* for 20 minutes. Free nuclei were concentrated by step gradient centrifugation and pelleted by standard centrifugation at 2,500 × *g* for 10 minutes. 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, following by digestion with proteinase K for 1 h at 50 °C. Unincorporated DL-Green label was removed from the reaction by absorption onto a

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 addition of Bionano DNA Stain at a final concentration of 1 microliter per 0.1 microgram 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 10X Chromium Sequencing Collection

DNA was isolated from approximately 50 g fast-frozen young leaf tissue using a protocol described by Luo and Wing, 2003. 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 2,500 × *g* for 15 minutes and washed with NIB+0.1% 2-mercaptoethanol followed by a second centrifugation at 2,500 × *g* for 15 minutes. The pellet was embedded in LMP agarose plugs, and 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

repair reaction. 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 each were used for Tx2783 and RTx436.

Chromium 10X 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 Mbp or 87× genome coverage. For RTx436, a total of 229,285,050 clusters were obtained, corresponding to 69.3 Mbp or 86× genome coverage.

Long- and Short-read Sequencing

Long-read data were generated using the Pacific BioSciences (Menlo Park, CA, USA) Sequel platform at Corteva Agriscience (Johnston, IA, USA). 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× genome coverage 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 10X Genomics (Pleasanton, CA, USA) Chromium libraries at Corteva Agriscience on the HiSeq 2500 platform (Illumina, San Diego, California). The Tx2783 Chromium dataset had 76× coverage and mean molecule length of 62,942bp. The coverage depths and mean molecule lengths for the Tx2783 and RTx436 Chromium libraries were 76×/62.9kb and 84×/98.5 kb, respectively.

Genome Assembly and Polishing

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Raw PacBio subreads were corrected and assembled using Canu v1.8 (Koren et al., 2017) 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 using samtools v1.9 (https://github.com/samtools/samtools) from the Long Ranger output BAM file.

Genome Mapping

Genome maps for Tx2783 and RTx436 were generating on the Bionano Genomics (San Diego, California) Saphyr platform at Corteva Agriscience 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_08222018) 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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

For RTx436, DLE-1 labeled molecule data from two flow cells (1 chip) were 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_10252018) 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 mis-assemblies in the contigs were identified and cut to resolve conflicts relative to the genome maps. In addition to auto-conflict 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 contigs 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 (Li, 2018). 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 to 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.

Manual Curation of the SCA Locus

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Visual inspection of the hybrid scaffold at the SCA locus revealed an approximately 69.8 kb collapsed repeat in the contig assembly that was not resolved during the initial manual curation process. In addition, alignment of the PacBio corrected reads to the contig assembly similarly showed evidence of a collapsed assembly of approximately 77.7 kb within the QTL locus (i.e. 2× corrected read coverage, soft-clipped reads flanking the potential repeat). The longest corrected reads from a 592-kb target region spanning the QTL region were manually assembled using GeneCodes Sequencher (Ann Arbor, Michigan) to generate a new sequence that resolved the collapsed repeat such that it was congruent to the Bionano map and incorporated the previously soft-clipped reads. As a result, 70,267 bp were added to the SCA locus. The entire manually curated region was swapped back into the pseudomolecule assembly and subjected to an additional round of sequence polishing with PacBio raw reads and 10X Chromium as described above. This was necessary because the corrected reads were used for the targeted assembly. Finally, the curated and polished 592-kb region was swapped back into the original pseudomolecule assembly. Thus, the final assembly, including the manually curated SCA locus was only subjected to a single round of polishing.

Annotation of the Tx2783 and RTx436 Assembly

Gene annotations were performed using a strategy that combined evidence-based and *ab initio* gene predictions. For evidence-based predictions, genome-guided transcript assemblies were generated from four different assemblers, Trinity (v2.8.4) (Grabherr et al., 2011), StringTie (v1.3.5) (Pertea et al., 2015), Cufflinks (v2.1.1) (Trapnell et al., 2010), and PsiCLASS (v1.0.1) (Song et al., 2019), using the transcriptome data from seven tissue types across the juvenile, vegetative, and reproductive stages of development (Supplemental Data S7). The best sets of transcripts were identified and annotated as genes using Mikado (v2.0rc6) (Venturini et al., 2018). Briefly, the RNA-seq reads from each library were mapped to STAR (v2.7.0d) (Dobin et al., 2013)-indexed Tx2783 and RTx436 genomes using a two-pass mapping approach. Default options were used for mapping with few post-processing options enabled. Individually mapped RNA-seq libraries were then pooled, sorted, and indexed using Picard (v2.7.0) (http://broadinstitute.github.io/picard) for transcript assembly programs. For all genome-

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

guided transcriptome assemblers, default options were used; if RNA-seq strandedness was allowed, it was set to stranded. For Trinity, maximum intron size was set to 10000. All assemblers generated a GFF3 as the final output except for Trinity, for which assemblies were mapped against the genome using GMAP (v2019-03-15) (Wu et al., 2005). Portcullis (v1.1.2) (Mapleson et al., 2018) was used to generate high-quality splice junctions from the merged mapped reads. Mikado was configured to use all transcript assemblies, portcullis-generated splice sites, and a plants.yaml scoring matrix. Preliminary transcripts were prepared for Mikado by 1) merging all transcripts and removing the redundant copies, 2) processed using TransDecoder (v5.5.0) (Grabherr et al., 2011) (to identify open reading frames), and 3) aligned with BLASTX (v2.2.29+) (Altschul et al., 1997) 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. These were provided as input for Mikado for picking and annotating the best transcripts for each locus. The resultant GFF3 file was used to extract transcripts and proteins using the gffread utility from the Cufflinks package.

Additional structural improvements to Mikado-generated transcripts were completed using the PASA (v2.3.3) (Haas et al., 2003) genome annotation tool. The inputs for PASA included 209,835 maize EST derived from GenBank, Mikado-assembled transcripts for Tx2783 and RTx436, 37,655 sorghum iso-seq transcripts from 11 developmental tissues, and 31,881 sorghum full-length cDNAs from RIKEN that were filtered for intron retention (Reddy et al., 2016; Wang et al., 2018). PASA was run with default options; in the first step, transcript evidence was aligned to the masked sorghum genomes using GMAP (v2019-03-15) (Wu et al., 2005) and Blat (v.36) (Kent, 2002). The full-length cDNA and Iso-seq transcript IDs were passed in a text file (-f FL.acc.list) during the PASA alignment step. Valid near-perfect alignments with 95% identity were clustered based on genome mapping location and assembled into gene structures that included the maximal number of compatible transcript alignments. PASA assemblies were then compared with sorghum Mikado transcript models using default parameters. PASA updated the models, providing UTR extensions, as well as novel and additional

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

alternative isoforms. PASA-generated models were passed through the MAKER-P (v3.0) (Campbell et al., 2014) annotation pipeline as model_gff along with all the transcript and protein sequences to yield Annotation Edit Distance (AED) scores to assess the quality of annotations. Transposon element (TE)-related genes were filtered using the TEsorter tool (Zhang et al., 2019), which uses the REXdb (viridiplantae_v3.0 + metazoa_v3) (Neumann et al., 2019) 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 (Hoff et al., 2015). Non-overlapping BRAKER gene models were updated with PASA, filtered for TE-related genes, provided with AED scores using MAKER, 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. Functional annotation was performed using InterProscan-5.38-76.0 (Jones et al., 2014). Canonical assignment of transcripts was performed using a combination of criteria including domain coverage, translation length, and AED scores comparing each transcript to sample specific transcripts assembled with Stringtie. Transcripts were ranked by each criterion, and a canonical transcript was selected by a ranked-choice voting scheme (Olson et al., 2020). The functional and canonical attributes were added to the annotation files, and quality checking was performed using GFF3toolkit (Chen et al., 2019).

Rampage Library Construction, Data Generation and Analysis

This protocol is a modified version of a previously published method (Batut et al., 2013). Prior to incubation with Terminator™ 5′-Phosphate-Dependent Exonuclease (TEX) (Lucigen) to remove all residual RNAs containing 5’ monophosphate, as described in Batut et al,. we removed all ribosomal RNAs using the RiboMinus™ Plant Kit for RNA-Seq (Thermo Fisher Scientific).

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 Batut et al.

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 (Wang et al., 2018). The mapped reads from each tissue were clustered using Paraclu (Frith et al., 2008). 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 (Aaron et al., 2010) 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 (Ramirez et al., 2016).

SV Detection

All available genomes were aligned to the sorghum Tx2783 genome using MUMmer 4 (Marcais et al., 2018) with default parameters. Plots of the alignment only used the one-to-one alignment. SNP and small indels (<50 bp) were also called by the tool "show-snp" in the MUMmer 4 package. Large indels and inversions were called using the SVMU pipeline (Chakraborty et al., 2019). Insertions and deletions (>50 bp) with any overlap with the gap region in the reference genome were filtered out. Functional annotation of

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SNPs was performed using SnpEff (Cingolani et al., 2012) based on the Tx2783 genome annotation. SIFT 4G (Vaser et al., 2016) was used to predict deleterious mutations through the SciApps platform (Wang et al., 2018).

Pan-Genome Construction and Resequencing Data Analysis

The pan-genome was constructed using Tx2783 as backbone, with the addition of the sequence of insertions which were called from the other four complete genomes. Resequencing data from Mace et al. 2013 and 17 BAP/SAP lines (https://datacommons.cyverse.org/browse/iplant/home/shared/terraref) were used for population analysis. Three lines (KS115; Yik.solate; Kilo) were removed due to low gene coverage; the remaining 61 genomes were mapped to this pan-genome using the BWA software (Li et al., 2009a). SAMtools (Li et al., 2009b) was used to convert mapping results to bam format. On one hand, BEDTools v2.29.2 (Aaron et al., 2010) was used to convert the bam files to bed format and calculate the gene coverage from the reads across 62 genomes. Next, core, softcore, dispensable and private genes were identified under three different cutoffs based on coverage: 50%; 75% and 90%; On the other hand, bam alignments were converted to pileup format using the pileup command. SNPs were called using both minimum base quality threshold and minimum mapping quality of 20. We then used vcf2phylip v2.0 (Ortiz, 2019) to convert the VCF to Phylip format, and then the python script ascbias.py (https://github.com/btmartin721/raxml_ascbias) was used to remove sites considered as invariable by RAxML. Finally, RAxML-NG (Kozlov et al., 2019) was used to estimate the best ML tree, then the bootstraps were used to plot support on the best tree.

Identification of NLR Loci and Gene Expression Analysis

Genome-wide nucleotide-binding and leucine-rich repeat (NLR) loci were annotated using NLR-Annotator (Steuernagel et al., 2020), and loci that overlapped with genes were extracted using BEDTools (Quinlan et al., 2010). Phylogenetic trees were constructed using the Interactive Tree Of Life (iTOL) (Letunic et al., 2019). Illumina raw

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

reads from Tetreault et al (Tetreault et al., 2019) were aligned to the Tx2783 reference genome using STAR 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. A k-means approach was used [R Bioconductor package 'Mfuzz' (Kumar et al., 2007)] to cluster dynamically expressed genes based on their expression profiles across different time points after sugarcane aphid infestation.

Identification of the SCA Resistance Locus

To identify the SCA resistance locus, the two parental lines Tx2783 and BTx623 and two pools from a recombinant inbred line (RIL) population, one consisting of 20 SCA-resistant individuals and the other consisting of 20 SCA-sensitive individuals, were sequenced on the Illumina platform. The average sequencing depth was about 10×. Bowtie2 (Langmead et al., 2012) was used to align the reads to the sorghum BTx623 v3.1.1 reference genome in '--very-sensitive' mode. Duplicates from sequencing were masked using Picard tools (http://broadinstitute.github.io/picard/). After alignment, variations in the four samples were called by samtools (Li et al., 2009), using only reads with sequencing and alignment quality above 20. Only sites sequenced in all four samples were considered for the subsequent analysis. Candidate variations were required to meet the following criteria: 1) homozygous in all four samples; 2) different genotypes in the two parental lines; 3) BTx623 and Tx2783 genotype in the sensitive and resistant pools, respectively. The effects of the remaining variants were then predicted by SNPeff (Cingolani et al., 2012). Only variations annotated with MISSENSE, NONSENSE, or SPLICE_SITE_ACCEPTOR/DONOR effects were considered to be large-effect variations. The variations remaining as candidates were counted in 100-kb sliding windows with a step size of 10 kb.

ACCESSION NUMBERS

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 ERZ1345851 and ERZ1345852, respectively. RIL population sequencing data are available under accession number PRJEB39196. RAMPAGE datasets are available under accession number PRJEB42222.

Figure Legends

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 are occurred at the centromeres. Chr00: Contigs unplaced into the chromosome. ChrC: Chlotoplast sequence. ChrM: Mitochondria sequence. **(B)**. BUSCO assessment of the indicated sorghum genomes.

Figure 2. RAMPAGE signal enrichment around TSSs in Tx2783 (A) and RTx436 (B) in root and shoot tissues.

Figure 3. 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 platinum sequence; **(E)** BTx623 Bionano map and RTx436 platinum sequence; **(F)** Tx2783 platinum sequence and RTx436 Bionano map.

Figure 4. Characterization of SNPs in sorghum genomes. (A) Genome-wide distribution of SNPs in each genome relative to the Tx2783 reference. **(B)** Distribution of SNPs in the indicated gene features. **(C)** Venn diagram of deleterious mutations in each

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

genome, using Tx2783 as a reference. **(D)** Gene Ontology enrichment of SNP-associated genes.

Figure 5. Characterization of large structural variations in sorghum genomes. (A) Circular visualization of different categories of SVs at the genome-wide level. a. Heatmap of gene density in the Tx2783 genome. b. Genome-wide distribution of INS density in four genomes: BTx623, RTx436, RTx430, and Rio. c. Genome-wide distribution of DEL density in four genomes: BTx623, RTx436, RTx430 and Rio. d. Genome-wide distribution of CNV density in four genomes: BTx623, RTx436, RTx430 and Rio. e. Genome-wide distribution of INV density in four genomes: BTx623, RTx436, RTx430 and Rio. **(B–E)** Overlap of SVs among sorghum genomes. **(F)** Distribution of INS SVs in different genomic regions. **(G)** Distribution of DEL SVs in different genomic regions.

Figure 6. Phylogenetic tree of 62 sorghum accessions and distribution of core genes identified from the population. (A) Phylogenetic tree of 62 genome accessions. Blue: Five complete genome assemblies; Purple: 17 BAP/SAP lines; Black: Lines from Mace et al., 2013. **(B)** Cumulative curve of covered protein-coding genes in 62 genomes under 75% gene coverage cutoff. **(C)** Distribution of core, softcore, dispensable, and private protein-coding genes identified from 62 sorghum accessions using the 75% gene coverage cutoff.

Figure 7. Identification and characterization of sugarcane aphid resistance locus. (A) Genome-wide distribution of NLR loci in each genome. **(B)** Phenotype of SCA resistance in sorghum line Tx2783. **(C)** Region of SCA resistance genes, identified using a RIL population. **(D)** Heatmap of genes within a 580-kb window of the SCA region across 62 sorghum accessions from resequencing data. Resistance genes are outlined in red on the left of the heatmap: SbiRTX2783.06G018200; SbiRTX2783.06G018300; SbiRTX2783.06G018400; SbiRTX2783.06G018450; SbiRTX2783.06G018460; SbiRTX2783.06G018500; SbiRTX2783.06G018600; SbiRTX2783.06G018700; SbiRTX2783.06G018800; SbiRTX2783.06G018900. **(E)** Expression of five genes in the

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SCA region from RNA-seq data. SbiRTX2783.06G018300; SbiRTX2783.06G018400; SbiRTX2783.06G018500; SbiRTX2783.06G018800; SbiRTX2783.06G018900.

Figure 8. Gene expression analysis after sugarcane aphid infestation. (A) Heatmap of genes expressed before and after SCA infestation. **(B)** Clusters of genes expressed before and after SCA infestation. **(C)** Heatmap of NLR genes before and after SCA infestation.

Table 1. **Genome assembly statistics for Tx2783 and RTx436.**

SUPPLEMENTAL INFORMATION

Supplemental Fig. S1 Number of different types of SVs and the number of overlapping genes in the indicated genomes: (**A–B**). BTx623; (**C–D**). Rio genome; (**E–F**). RTx436 genome; (**G–H**). RTx430 genome.

Supplemental Fig. S2 Correlation of number of different SVs with the number of SNPs in each sorghum genome.

Supplemental Fig. S3 (**A**). Distribution of number of CNVs per gene in each genome. (**B**). Distribution of size of CNVs in each genome.

Supplemental Fig. S4 Correlation of number of CNVs with Size of CNVs in each sorghum genome.

Supplemental Fig. S5 Gene Ontology enrichment of genes overlapping various SVs: (**A**) Insertion. (**B**) Deletion. (**C**). CNV.

Supplemental Fig. S6 Characterization of core non-coding genes. (A) Cumulative curve of non-coding genes across 62 sorghum accessions. (B) Distribution of core, softcore, dispensable and private non-coding genes identified from 62 sorghum accessions.

Supplemental Fig. S7 Chromosome distribution of NBS-LRR genes in sorghum (**A**) and maize (**B**).

Supplemental Fig. S8 Manual curation of assembly in the SCA region.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplemental Fig. S9 Dotplot of deletions in the SCA resistance region between Tx2783 and BTx623.

Supplemental Table S1. Summary of Bionano data for Tx2783, RTx436 and BTx623.

Supplemental Table S2. Summary of TE annotations for Tx2783, RTx436, and BTx623 genomes from EDTA.

Supplemental Table S3. Status of protein-coding gene annotations for Tx2783 and RTx436.

Supplemental Table S4. The list of large inversions that detected in Tx2783 and RTx436.

Supplemental Table S5. Number of SNPs called in each genome relative to the Tx2783 reference.

Supplemental Table S6. Number of SVs called in each genome relative to the Tx2783 reference.

Supplemental Table S7. Number of core, softcore, dispensable and private genes under various cutoffs of gene coverage for protein-coding and non-coding genes identified from 62 sorghum accessions.

Supplemental Table S8. Number of NLR loci in each genome, identified by genome scanning.

Supplemental Table S9. Summary of SVs in the SCA resistance region.

Supplemental Data S1. Structural variations called from Bionano genome maps between different genomes.

Supplemental Data S2. Structural variations between different genomes called from whole-genome sequence alignments using Tx2783 as reference.

Supplemental Data S3. Information of germplasm that was used for this study.

Supplemental Data S4. Mapping status of different sorghum accessions to the pan genome.

Supplemental Data S5. Large-effect SNPs in BTx623 and Tx2783 at SCA region.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplemental Data S6. InterProScan result of genes within the SCA region.

Supplemental Data S7. RNA-seq data that is used for genome annotation.

AUTHOR CONTRIBUTIONS

B.W., Y.J., and D.W. conceived the idea for the study; V.L and K.F. generated genome assemblies; K.C., A.O. B.W. XH.W., Y.J. and XF.W generated genome annotations; M.R., T.G, JD. generated RNA, RNA-seq and RAMPAGE libraries; B.W analyzed the SNP and SV: B.W, L.W and A.O population analyses and core and dispensable gene. C.H. and Z.X. provided SCA seeds; S.A. performed the test of Peterson's sorghum lines for sugarcane aphid resistance; J.H. developed and selected the SCA-resistance RIL population. Y.H. constructed the sugarcane aphid resistance RIL population; Y.J. mapped the sugarcane aphid resistance region; B.W., Y.J., and D.W. wrote the manuscript.

ACKNOWLEDGEMENTS

We thank John Burke, Geoffrey Morris, Bob Klein, Bill Rooney, and Gary Peterson for helpful input concerning germplasm selection. This work was supported by USDA-SCA grant #51530311 and NSF Gramene grant #52930511. The authors thank Peter Van Buren for help with the computational system.

COMPETING INTERESTS

The authors declare that they have no competing interests.

REFERENCES

Alkan C, Coe BP, Eichler EE. (2011). Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* 12: 363–376.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 25(17)3389–3402.

Baker M. (2012). Structural variation: the genome's hidden architecture. Nat. Methods. 9: 133–137.

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

Batut P, Gingeras TR. (2013). RAMPAGE: promoter activity profiling by paired-end sequencing of 5'-complete cDNAs. Curr Protoc Mol Biol. 104:Unit 25B.11. doi: 10.1002/0471142727.mb25b11s104.

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

Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, Lei JK, Achawanantakun R, Jiao D, Lawrence CJ, 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.

Chen MM, Lin H, Chiang LM, Chilbers CP, Poelchau MF. (2019). The GFF3toolkit: QC and Merge Pipeline for Genome Annotation. Methods Mol Biol. 1858:75-87.

Chakraborty M, Emerson J, Macdonald S, Long A. (2019). Structural Variants Exhibit Widespread Allelic Heterogeneity and Shape Variation in Complex Traits. Nat Commun. 10(1):4872.

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

Cooper EA, Brenton ZW, Flinn BS, Jenkins J, Shu S, Flowers D, Luo F, Wang Y, Xia P, Barry K, et al. (2019). A New Reference Genome for Sorghum Bicolor Reveals High Levels of Sequence Similarity Between Sweet and Grain

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Genotypes: Implications for the Genetics of Sugar Metabolism. BMC Genomics. 20(1):420.

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

Dobin A, Davis CA, Schlesinger F, Drenkow, J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29:15–21.

Frith M, Valen E, Krogh A, Hayashizaki Y, Carninci P, Sandelin A. (2008). A code for transcription initiation in mammalian genomes. Genome Res. 2008 Jan;18(1):1-12. doi: 10.1101/gr.6831208.

Fuentes RR, Chebotarov D, Duitama J, Smith S, De la Hoz JF, Mohiyuddin M, Wing RA, McNally KL, Tatarinova T, Grigoriev A, et al. (2019). Structural variants in 3000 rice genomes. Genome Res. 29(5):870-880.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, 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(7):644-52.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, 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. Nature Biotechnology. 29(7)644–652.

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith JRK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, et al. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*. **31**, 5654–5666.

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. (2015). BRAKER1: Unsupervised Rna-Seq-Based Genome Annotation with Genemark-et and Augustus. *Bioinformatics*. 32 (5): 767-9.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Horiguchi G, Gonzalez N, Beemster GT, Inzé D, Tsukaya H. (2009). Impact of segmental chromosomal duplications on leaf size in the grandifolia-D mutants of Arabidopsis thaliana. Plant J. 60: 122–133.

Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin CS, et al. (2017). Improved maize reference genome with single-molecule technologies. Nature. 546(7659):524-527.

Jones P, Binns D, Chang HY, Fraser M, Li WZ, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. (2014). InterProScan 5: Genome-Scale Protein Function Classification. Bioinformatics. 30(9):1236-40.

Kent WJ. (2002). BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656–664.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research. 27(5):722-736.

Kozlov A, Darriba D, Flouri T, Morel B, Stamatakis A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 35(21):4453-4455. doi: 10.1093/bioinformatics/btz305.

Kumar L, Futschik ME. (2007). Mfuzz: A software package for soft clustering of microarray data. Bioinformation. 2(1): 5–7.

Langmead B, Salzberg S. (2012). Fast Gapped-Read Alignment With Bowtie2. Nat Methods. 9(4):357-9.

LeBauer, David et al. (2020), Data From: TERRA-REF, An open reference data set from high resolution genomics, phenomics, and imaging sensors, Dryad, Dataset, https://doi.org/10.5061/dryad.4b8gtht99.

Letunic I, Bork P. (2019). Interactive Tree Of Life (iTOL) v4: Recent Updates and New Developments. Nucleic Acids Res. 47(W1):W256-W259.

Luo M, Wing RA. (2003). An improved method for plant BAC library construction. Methods Mol. Biol. 236: 3–20.

Li H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 34:3094-3100.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Li H, Durbin R. (2009a). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 25(14): 1754–1760. doi: 10.1093/bioinformatics/btp324.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, et al. (2009b). The Sequence Alignment/Map Format and SAMtools. Bioinformatics. 25(16):2078-9.

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

Mace E, Tai SS, Innes D, Godwin I, Hu WS, Campbell B, Gilding E, Cruickshank A, Prentis P, Wang J, Jordan D. (2014). The plasticity of NBS resistance genes in sorghum is driven by multiple evolutionary processes. *BMC Plant Biol*. **14**:253.

Maron LG, Guimarães CT, Kirst M, Albert PS, Birchler JA, Bradbury PJ, Buckler ES, Coluccio AE, Danilova TV, Kudrna D, et al. (2013). Aluminum tolerance in maize is associated with higher MATE1 gene copy number. Proc. Natl. Acad. Sci. USA. 110: 5241–5246.

Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. (2019). Structural Variant Calling: The Long and the Short of It. Genome Biol. 20(1):246.

Mapleson D, Venturini L, Kaithakottil G, Swarbreck D. (2018). Efficient and accurate detection of splice junctions from RNA-seq with portcullis. GigaScience. 7:giy131.

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

McCormick RF, Truong SK, Sreedasyam A, Jenkins J, Shu SQ, Sims D, Kennedy M, Amirebrahimi M, Weers BD, 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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW. (2003). Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. The Plant cell. 15, 809–834.

Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR. (2018). High Contiguity Arabidopsis Thaliana Genome Assembly With a Single Nanopore Flow Cell. Nat Commun. 9(1):541.

Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ, Acharya CB, Mitchell SE, et al. (2013). Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc Natl Acad Sci U S A.* **110**:453-8.

Neumann P, Novák P, Hoštáková N, 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.

Olson A, Ware D. (2020). Ranked Choice Voting for Representative Transcripts with TRaCE. bioRxiv. doi: https://doi.org/10.1101/2020.12.15.422742.

Ou SJ, Su WJ, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, et al. (2019). Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline. Genome Biol. 20(1): 275.

Ordonio R, Ito Y, Morinaka Y, Sazuka T, Matsuoka M. (2016). "Chapter Five - molecular breeding of *Sorghum bicolor*, a novel energy crop," in *International Review of Cell and Molecular Biology*, ed. Jeon K. W. (Cambridge, MA: Academic Press), 221–257. 10.1016/bs.ircmb.2015.09.001.

Ortiz EM. 2019. vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis. DOI:10.5281/zenodo.2540861.

Paterson AH, Bowers JE, 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**:551-6.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology. 33(3):290-5.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Quinlan AR, Hall IM. (2010). BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features. Bioinformatics. 26(6):841-2.

Ramírez F, Ryan D, Grüning B, Bhardwaj V, Kilpert F, Richter A, Heyne S, Dündar F, Manke T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 44: W160–W165. doi: 10.1093/nar/gkw257.

Reddy AS, Abdel-Ghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, Schilkey F, Ben-Hur A, Reddy ASN. (2016). A survey of the sorghum transcriptome using single-molecule long reads. Nat Commun. 7:11706.

Saxena RK, Edwards D, Varshney RK. (2014). Structural variations in plant genomes. *Brief Funct Genomics*. **13**: 296-307.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al., (2009). The B73 maize genome: complexity, diversity, and dynamics. Science. 326(5956):1112-5. doi: 10.1126/science.1178534.

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, et al. (2007). Strong association of de novo copy number mutations with autism. *Science*. 316: 445–449.

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. (2018). Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 15(6):461-468.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 1;31(19):3210-2. doi: 10.1093/bioinformatics/btv351.

Soyk S, Lemmon ZH, Sedlazeck FJ, Jiménez-Gómez JM, Alonge M, Hutton SF, Van Eck J, Schatz MC, Lippman ZB. (2019). Duplication of a domestication locus neutralized a cryptic variant that caused a breeding barrier in tomato. Nat Plants. 5(5):471-479.

Song L, Sabunciyan S, Yang G, Florea L. (2019). A multi-sample approach increases the accuracy of transcript assembly. Nat Commun. 10, 5000.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Stefansson H, Rujescu D, Cichon S, Pietiläinen OP, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, et al. (2008). Large recurrent microdeletions associated with schizophrenia. *Nature.* 455: 232–236.

Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang C, Chougule K, Gao D, Iwata A, Goicoechea JL, et al. (2018) Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. Nat Genet. 50(11):1618. doi: 10.1038/s41588-018-0261-2. PMID: 30291357.

Steuernagel B, Witek K, Krattinger SG, Ramirez-Gonzalez RH, Schoonbeek HJ, Yu GT, Baggs E, Witek AI, Yadav I, Krasileva KV, et al. (2020). The NLR-Annotator Tool Enables Annotation of the Intracellular Immune Receptor Repertoire. Plant Physiol. 183(2):468-482.

Tello-Ruiz M, Naithani S, Gupta P, Olson A, Wei S, Preece J, Jiao Y, Wang B, Chougule K, Garg P, et al. (2020). Gramene 2021: harnessing the power of comparative genomics and pathways for plant research. Nucleic Acids Res. gkaa979. doi: 10.1093/nar/gkaa979. Online ahead of print.

Tetreault HM, Grover S, Scully ED, Gries T, Palmer NA, Sarath G, Louis J, Sattler SE. (2019). Global Responses of Resistant and Susceptible Sorghum (Sorghum bicolor) to Sugarcane Aphid (Melanaphis sacchari). *Front Plant Sci.* **10**:145.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Baren MJV, Salzberg SL, Wold BJ, Pachter L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology. 28(5):511-5.

Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. (2016). SIFT missense predictions for genomes. Nat. Protoc. 11: 1–9.

Venturini L, Caim S, Kaithakottil G, Mapleson DL, Swarbreck D. (2018). Leveraging multiple transcriptome assembly methods for improved gene structure annotation. GigaScience, 7,8. giy093.

Wang B, Regulski M, Tseng E, Olson A, Goodwin S, McCombie WR, Ware D. (2018). A comparative transcriptional landscape of maize and sorghum obtained by single-molecule sequencing. *Genome Res.* **28**, 921–932.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Wang L, Lu Z, Van Buren P, Ware D. (2018). SciApps: a cloud-based platform for reproducible bioinformatics workflows. Bioinformatics. 34: 3917–3920.

Wu TD, Watanabe CK. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 21(9):1859–1875.

Xiao H, Jiang N, Schaffner E, Stockinger EJ, van der Knaap E. (2008). A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. Science. 319: 1527–1530.

Ye J, Zhang Y, Cui HH, Liu JW, Wu YQ, Cheng Y, Xu HX, Huang XX, Li ST, Zhou A, et al. (2018). WEGO 2.0: a web tool for analyzing and plotting GO annotations, 2018 update. Nucleic Acids Res. 46: W71–W75.

Zarrei M, MacDonald JR, Merico D, Scherer SW. (2015). A copy number variation map of the human genome. *Nat Rev Genet*. **16**:172-83.

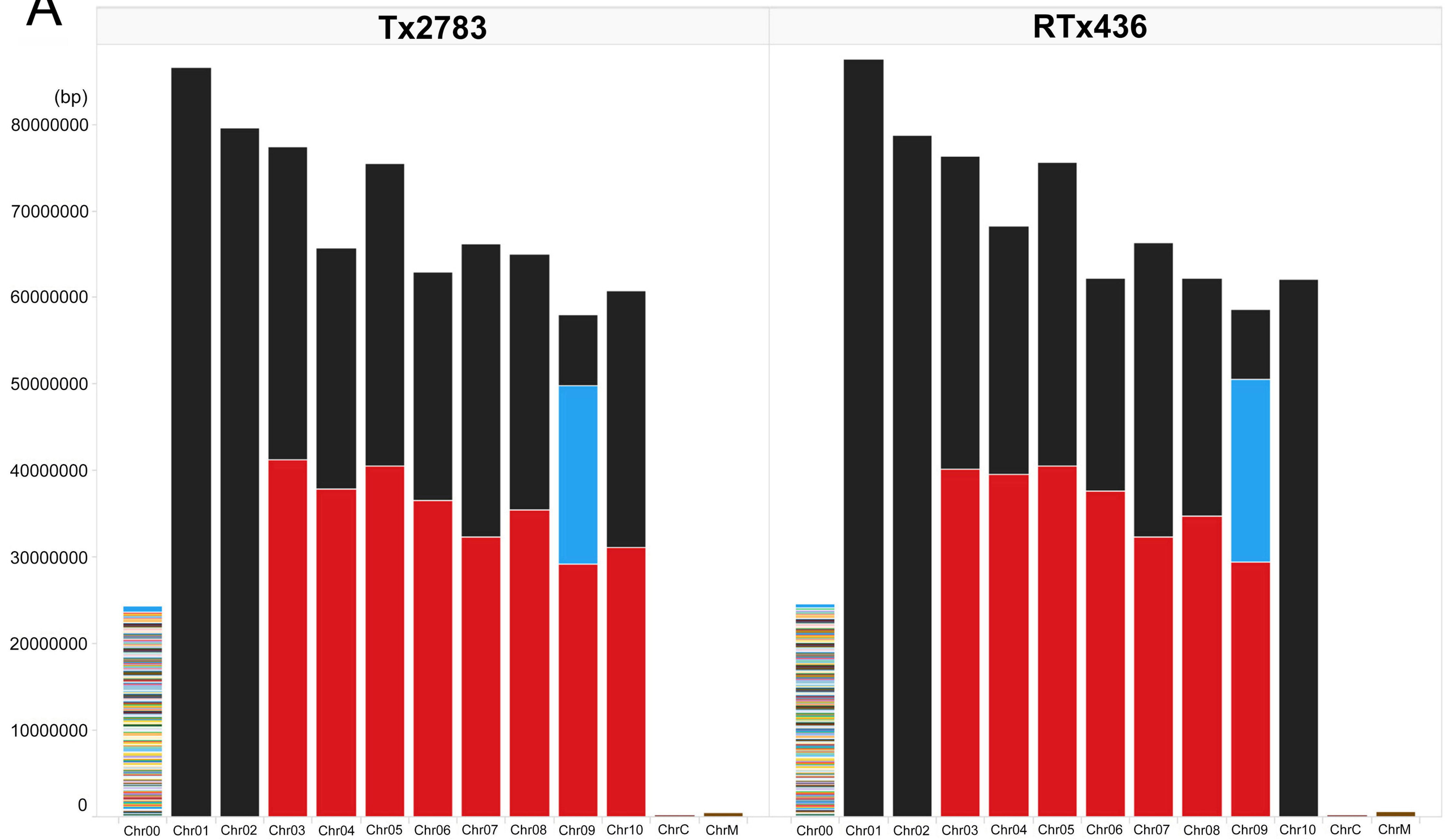
Zhang JW, Chen LL, Xing F, Kudrna DA, Yao W, Copetti D, Mu T, Li WM, Song JM, Xie WB, et al. (2016). Extensive sequence divergence between the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui 63. PNAS. 113 (35): E5163-E5171.

Zhang ZH, Mao LY, Chen HM, Bu FJ, Li GC, Sun JJ, Li S, Sun HH, Jiao C, Blakely R, et al. (2015). Genome-Wide Mapping of Structural Variations Reveals a Copy Number Variant That Determines Reproductive Morphology in Cucumber. Plant Cell. 27(6):1595-604.

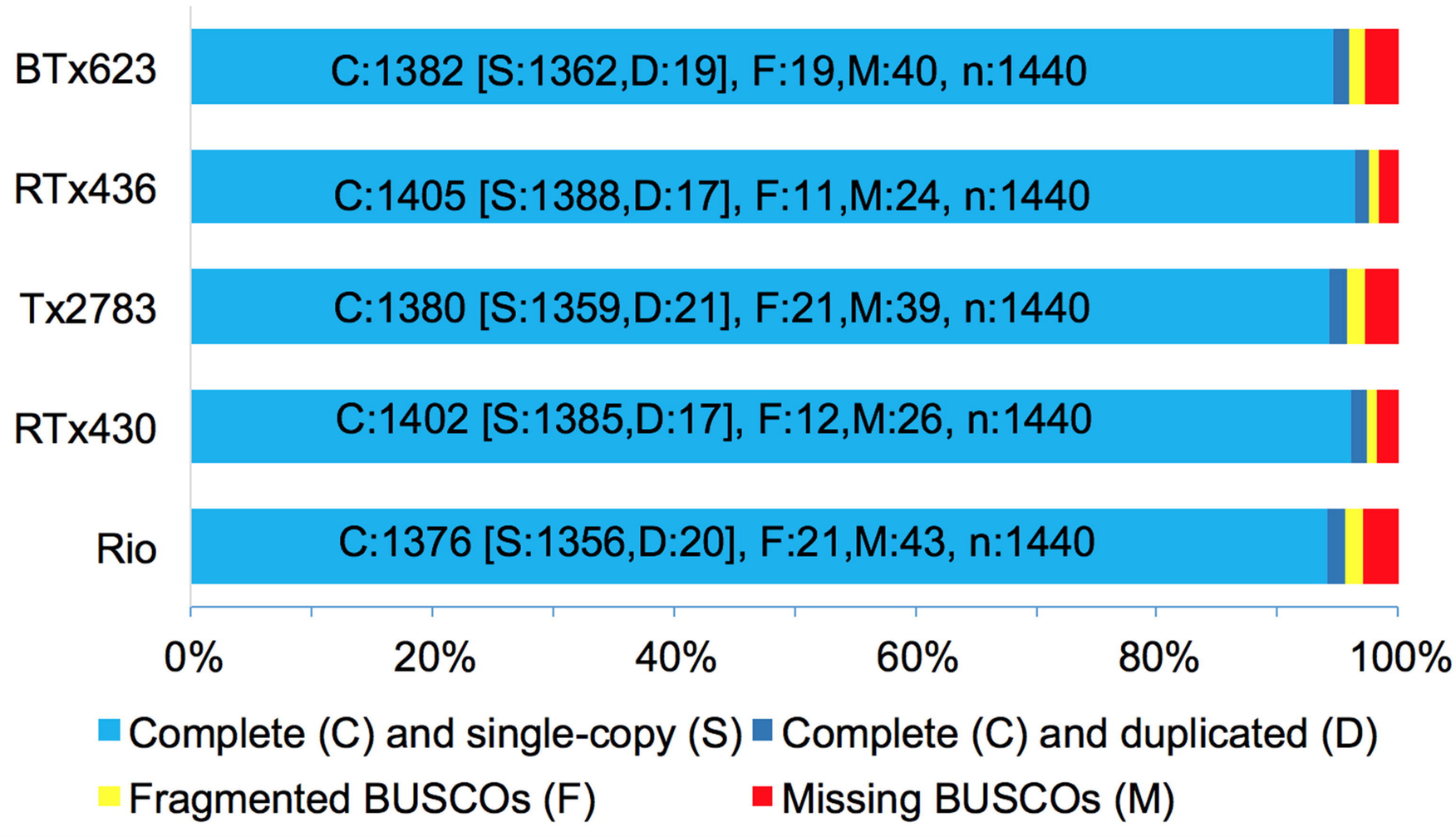
Zhang RG, Wang ZX, Ou S, Li GY. (2019). TEsorter: lineage-level classification of transposable elements using conserved protein domains. bioRxiv. doi:10.1101/800177.

Zhou T, Wang Y, Chen JQ, Araki H, Jing Z, Jiang K, Shen J, Tian D. (2004). Genome-wide identification of NBS genes in *japonica* rice reveals significant expansion of divergent non-TIR NBS-LRR genes. Mol Genet Genomics. 271, 402–415.

A

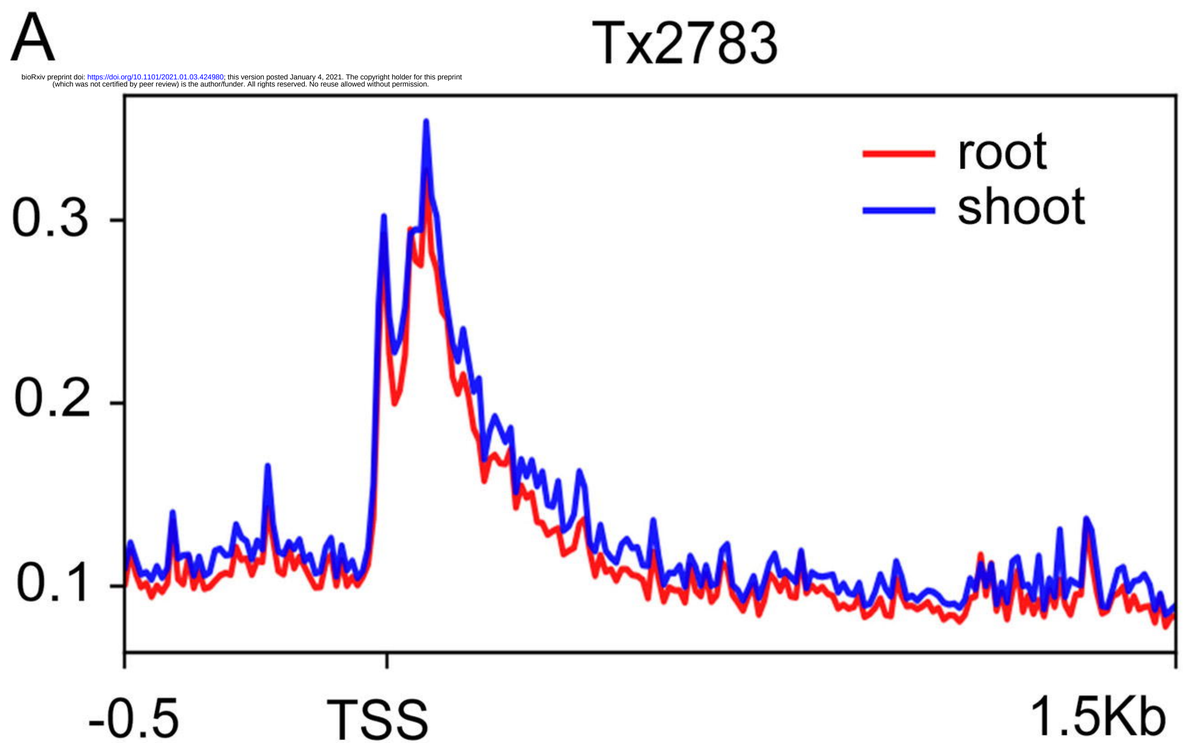


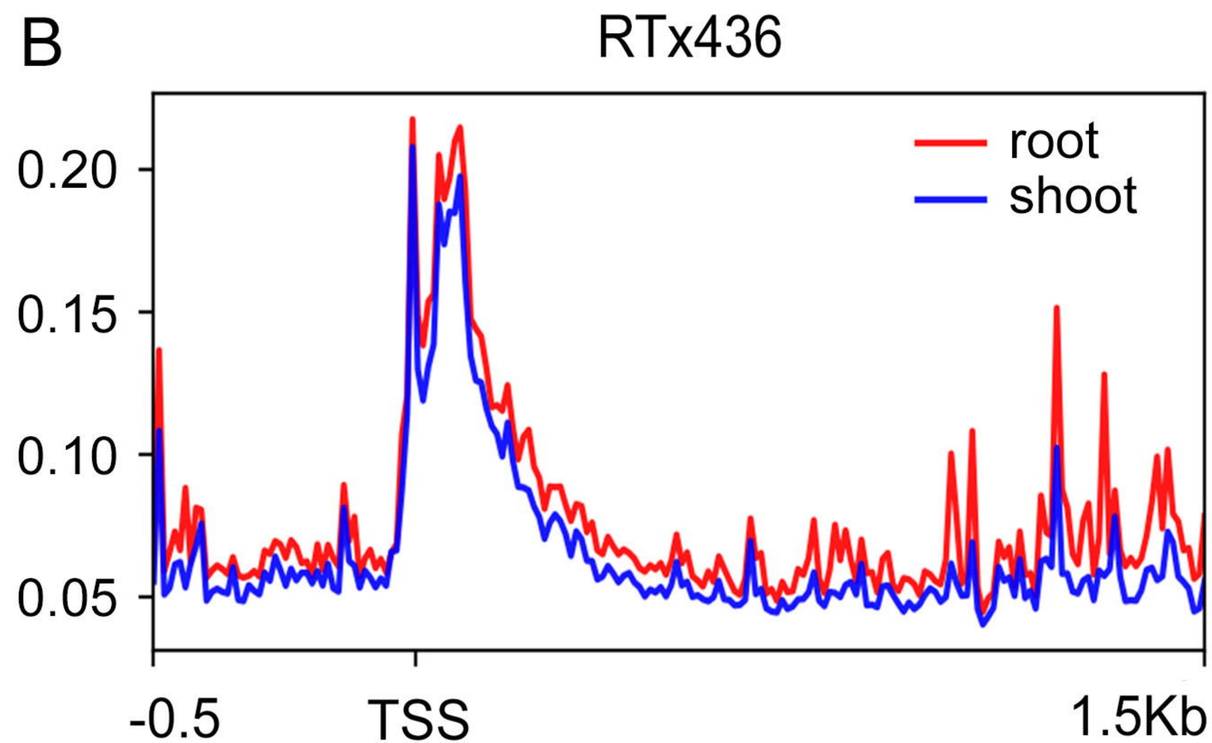
B

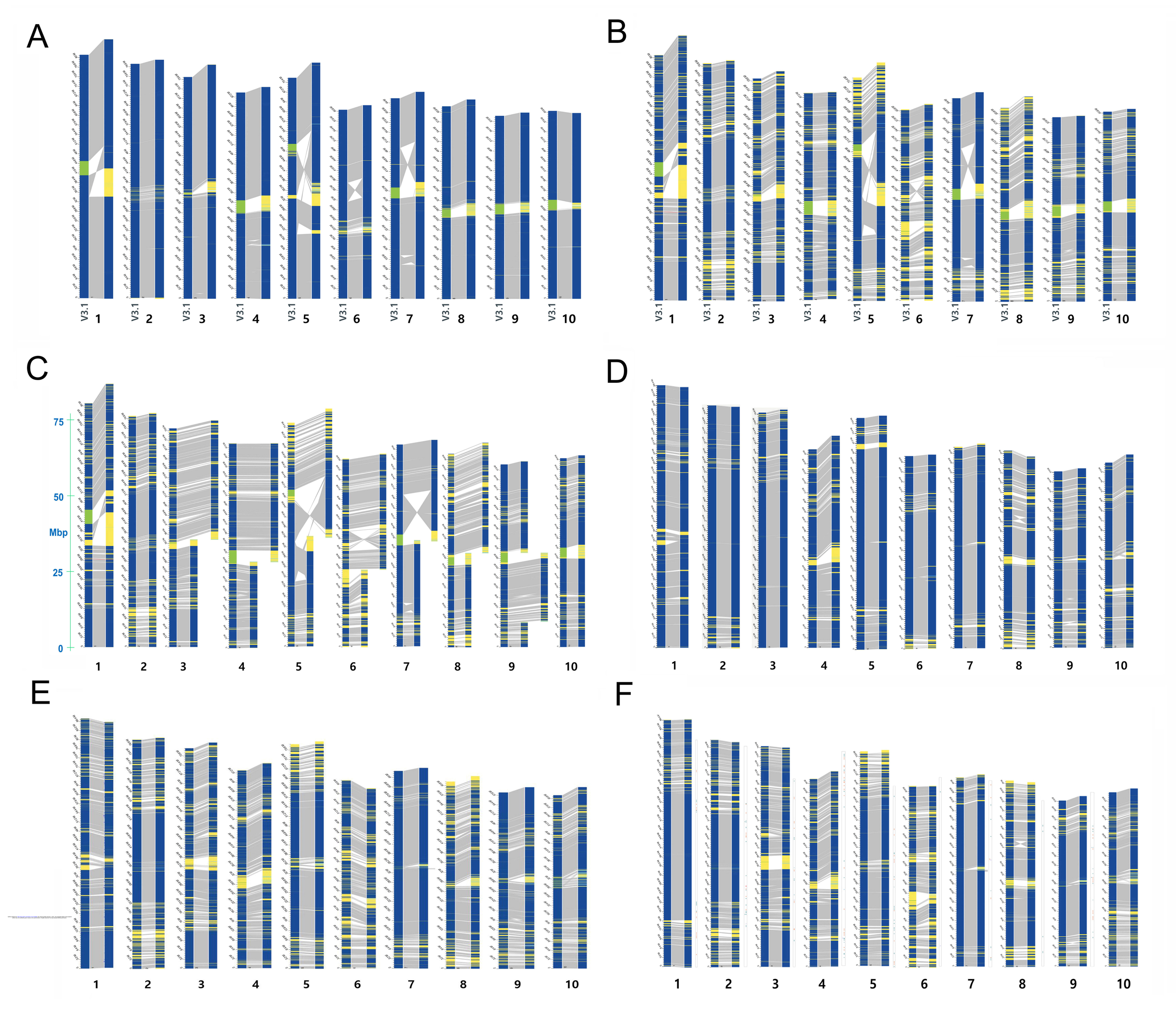


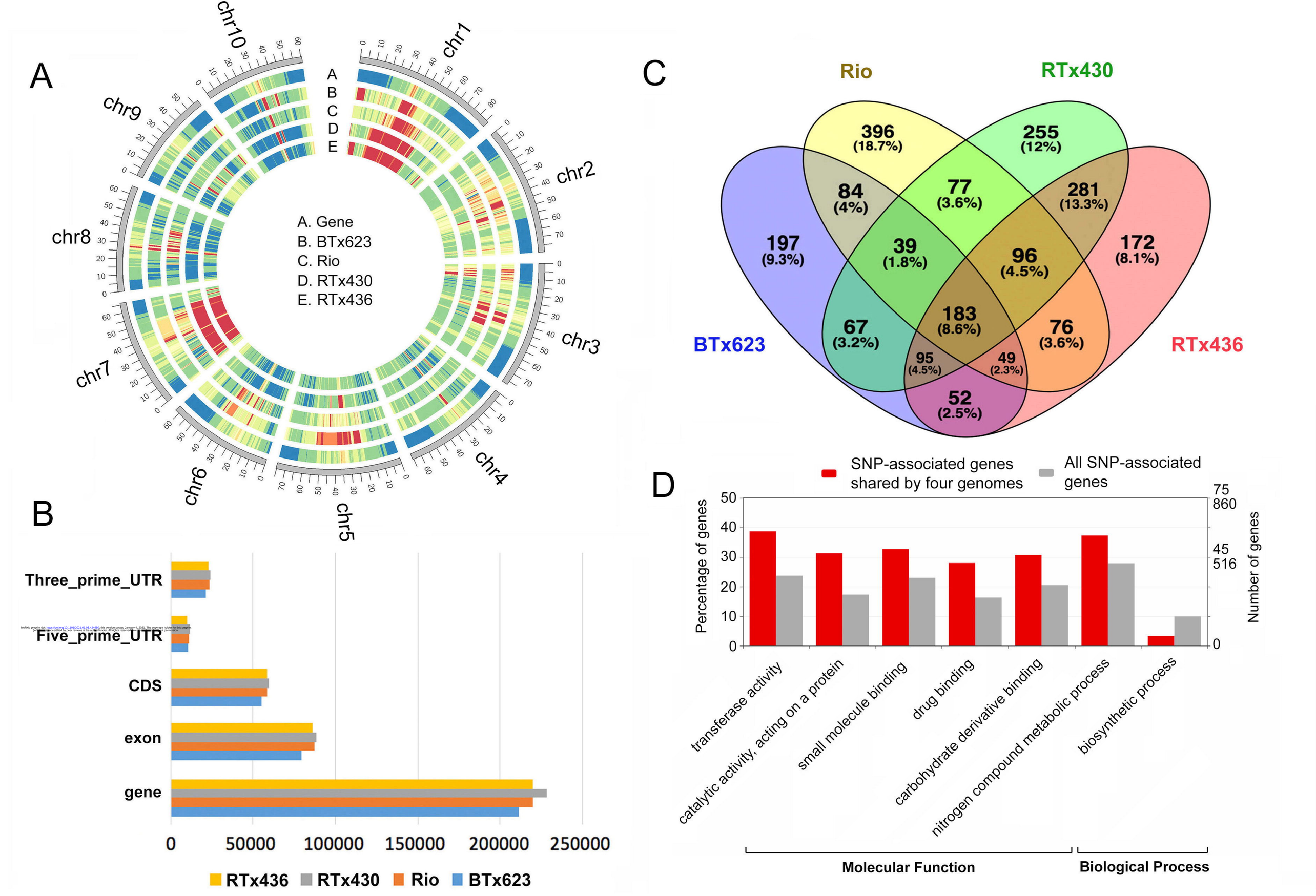
bioRxiv preprint doi: https://doi.org/10.1101/2021.01.02.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



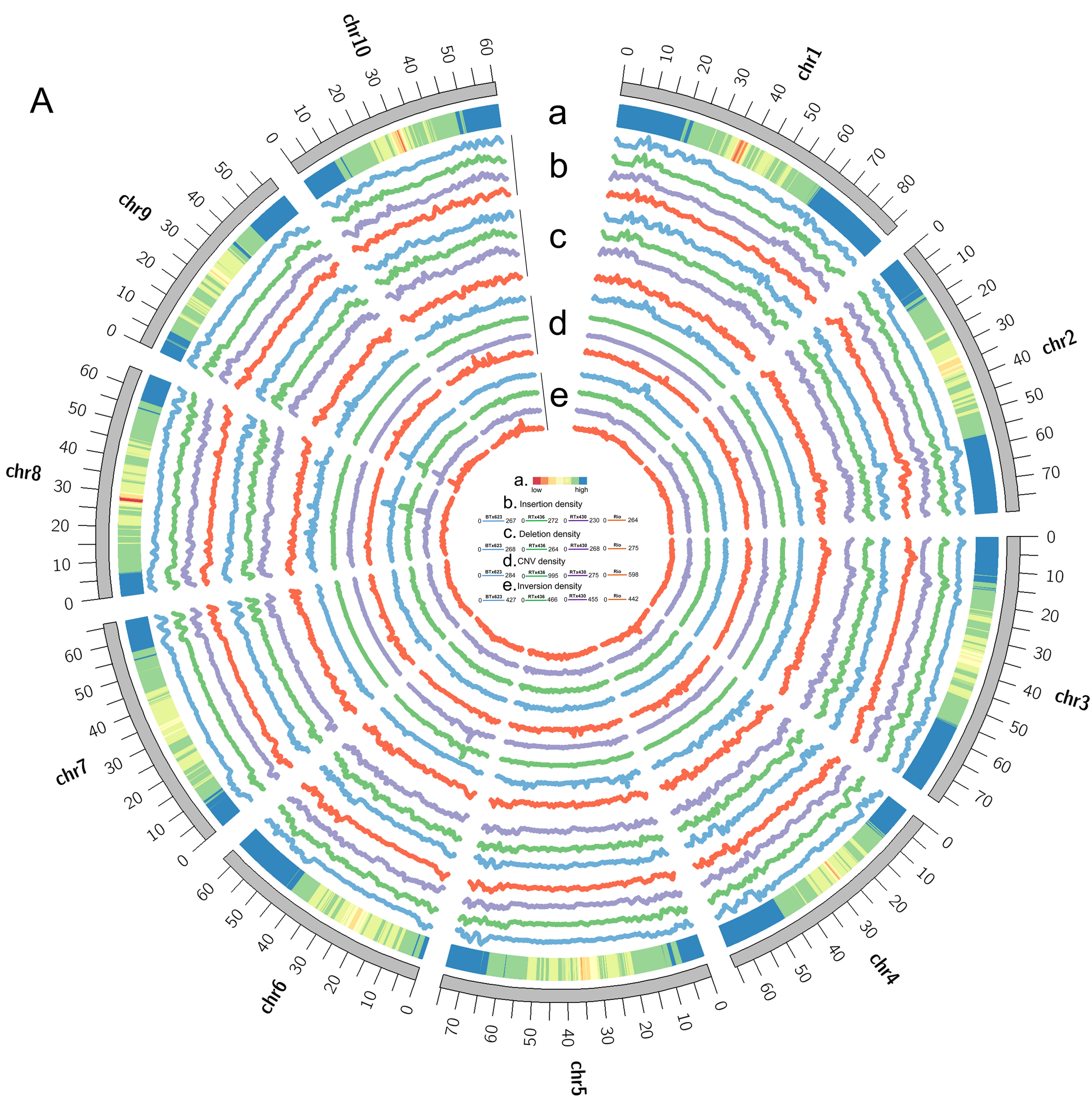



A
B
C
D
E
F
V3.1 1
V3.1 2
V3.1 3
V3.1 4
V3.1 5
V3.1 6
V3.1 7
V3.1 8
V3.1 9
V3.1 10
75
50
25
0
Mbp
1
2
3
4
5
6
7
8
9
10

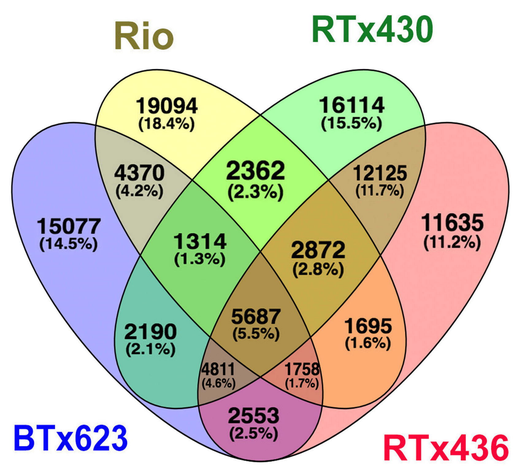


bioRxiv preprint doi: https://doi.org/10.1101/2021.01.03.424980; this version posted January 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

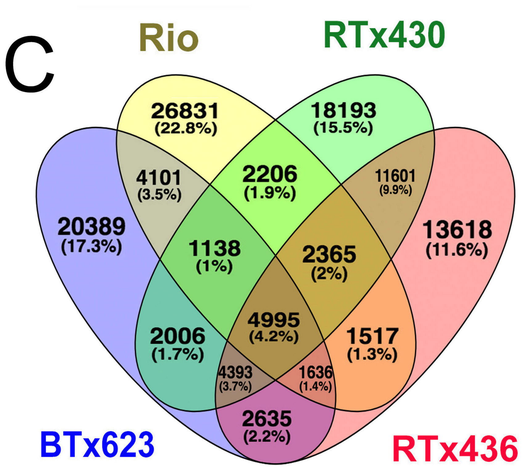
A


chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
a
b
c
d
e
a. low high
b. Insertion density
BTx623 267 RTx436 272 RTx430 230 Rio 264
c. Deletion density
BTx623 268 RTx436 264 RTx430 268 Rio 275
d. CNV density
BTx623 284 RTx436 995 RTx430 275 Rio 598
e. Inversion density
BTx623 427 RTx436 466 RTx430 455 Rio 442

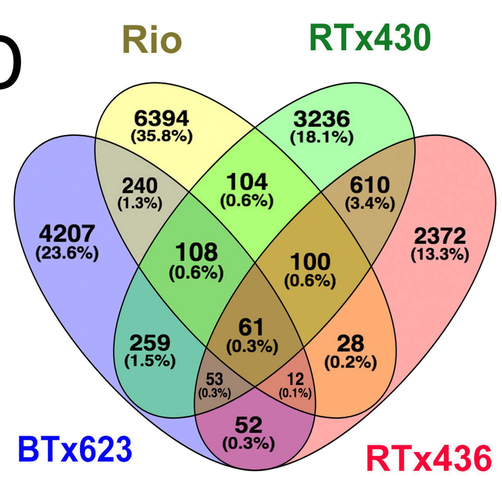
B


Rio
RTx430
BTx623
RTx436
19094 (18.4%)
16114 (15.5%)
15077 (14.5%)
11635 (11.2%)
4370 (4.2%)
2362 (2.3%)
12125 (11.7%)
1314 (1.3%)
2872 (2.8%)
2190 (2.1%)
5687 (5.5%)
1695 (1.6%)
4811 (4.6%)
1758 (1.7%)
2553 (2.5%)

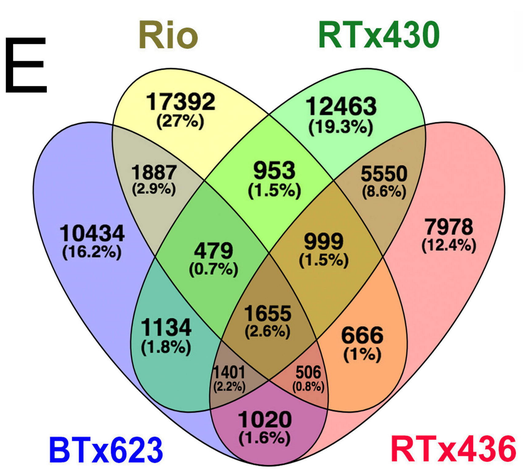
C


Rio
RTx430
BTx623
RTx436
26831 (22.8%)
18193 (15.5%)
20389 (17.3%)
13618 (11.6%)
4101 (3.5%)
2206 (1.9%)
11601 (9.9%)
1138 (1%)
2365 (2%)
2006 (1.7%)
4995 (4.2%)
1517 (1.3%)
4393 (3.7%)
1636 (1.4%)
2635 (2.2%)

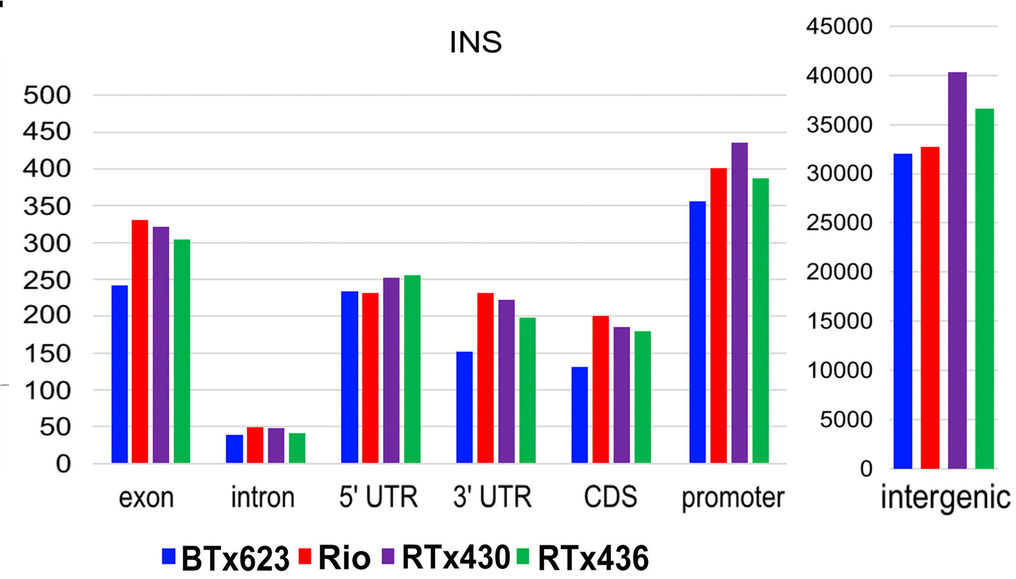
D


Rio
RTx430
BTx623
RTx436
6394 (35.8%)
3236 (18.1%)
4207 (23.6%)
2372 (13.3%)
240 (1.3%)
104 (0.6%)
610 (3.4%)
108 (0.6%)
100 (0.6%)
259 (1.5%)
61 (0.3%)
28 (0.2%)
53 (0.3%)
12 (0.1%)
52 (0.3%)

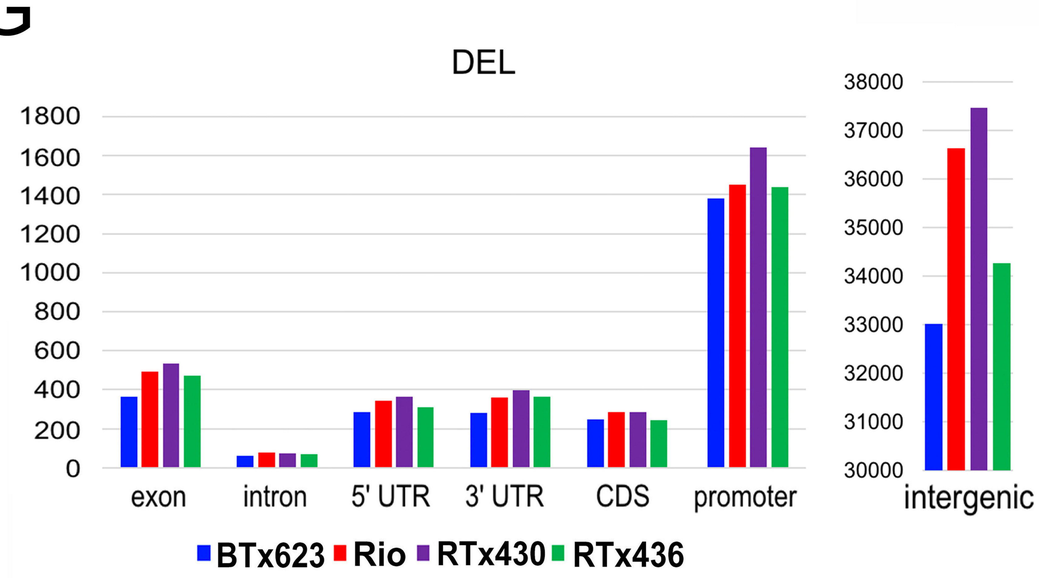
E

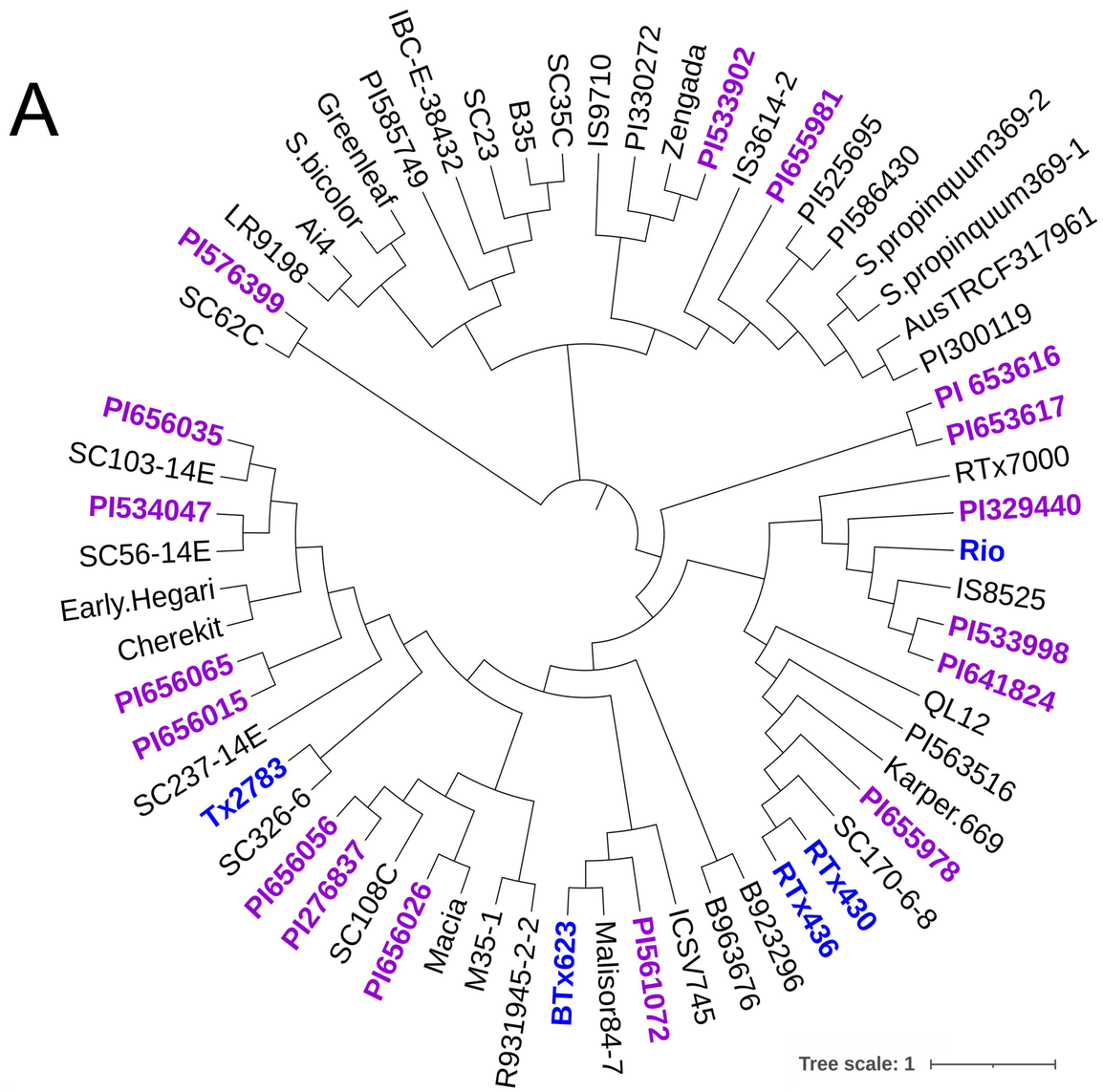

Rio
RTx430
BTx623
RTx436
17392 (27%)
12463 (19.3%)
10434 (16.2%)
7978 (12.4%)
1887 (2.9%)
953 (1.5%)
5550 (8.6%)
479 (0.7%)
999 (1.5%)
1134 (1.8%)
1655 (2.6%)
666 (1%)
1401 (2.2%)
506 (0.8%)
1020 (1.6%)

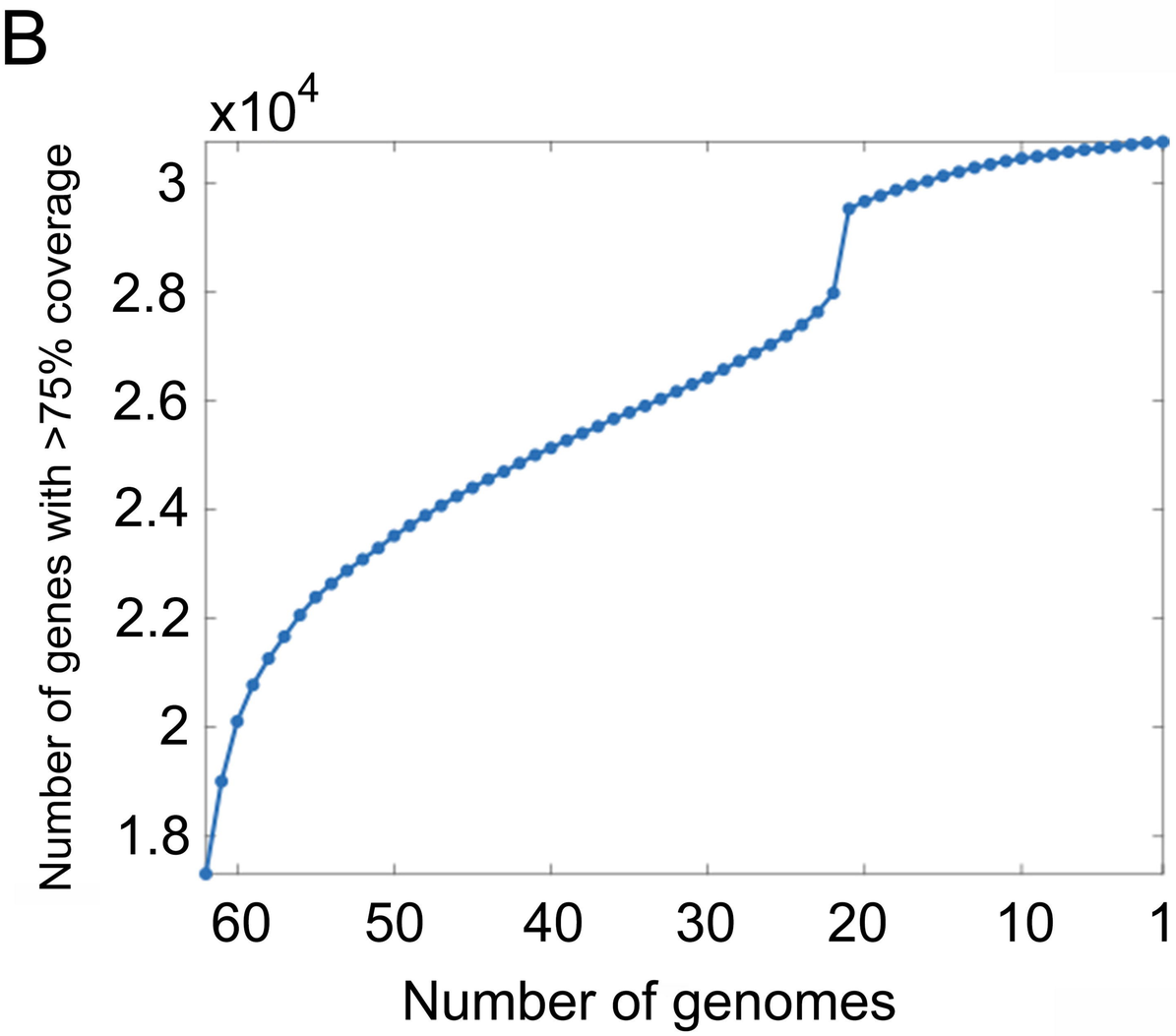
F

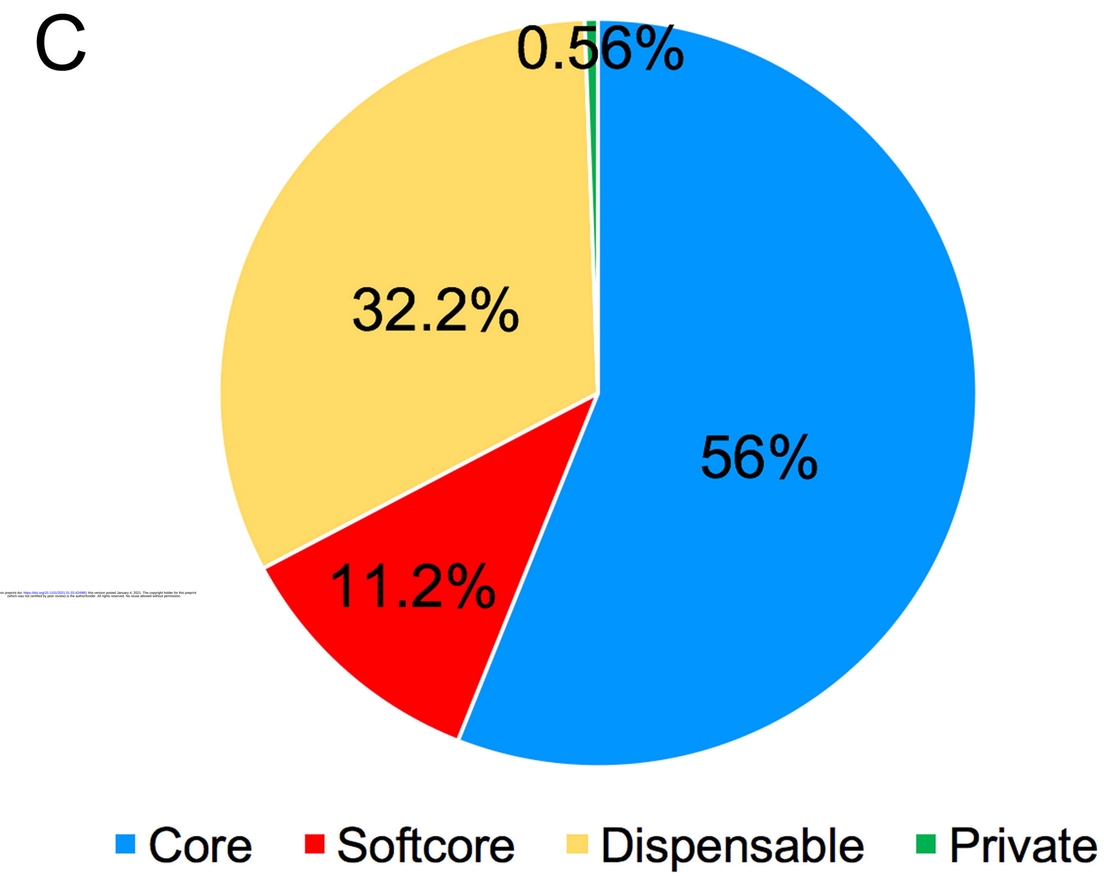

INS
exon
intron
5' UTR
3' UTR
CDS
promoter
intergenic
BTx623 Rio RTx430 RTx436

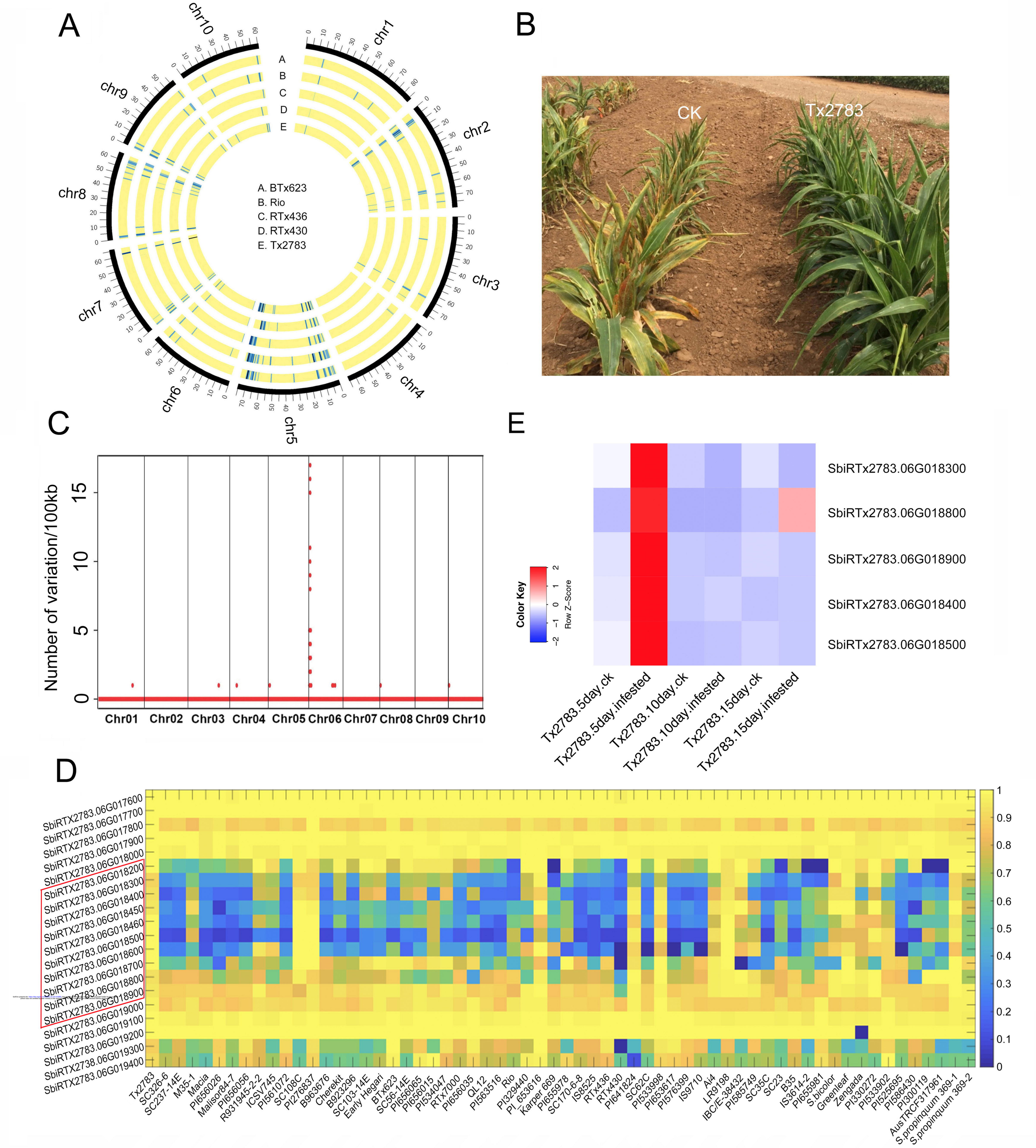
G

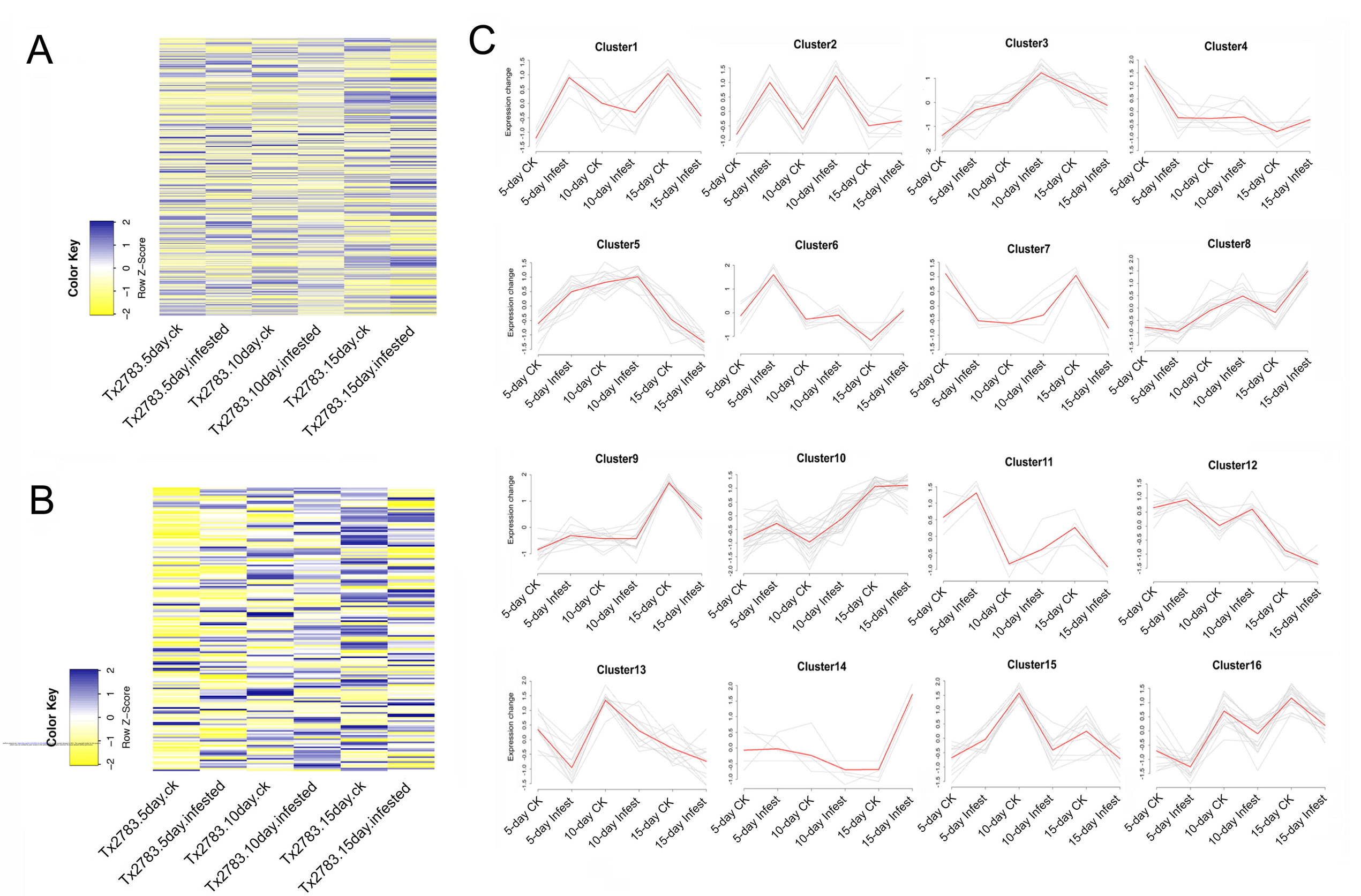

DEL
exon
intron
5' UTR
3' UTR
CDS
promoter
intergenic
BTx623 Rio RTx430 RTx436


A
IBC-E-38432
PI585749
Greenleaf
S.bicolor
Ai4
LR9198
PI576399
SC62C
SC23
B35
SC35C
IS9710
PI330272
Zengada
PI533902
IS3614-2
PI655981
PI525695
PI586430
S.propinquum369-2
S.propinquum369-1
AusTRCF317961
PI300119
PI 653616
PI653617
RTx7000
PI329440
Rio
IS8525
PI533998
PI641824
QL12
PI563516
Karper.669
PI655978
SC170-6-8
RTx430
RTx436
B923296
B963676
ICSV745
PI561072
Malisor84-7
BTx623
R931945-2-2
M35-1
Macia
PI656026
SC108C
PI276837
PI656056
SC326-6
Tx2783
SC237-14E
PI656015
PI656065
Cherekit
Early.Hegari
SC56-14E
PI534047
SC103-14E
PI656035
Tree scale: 1


B
x10^4
3
2.8
2.6
2.4
2.2
2
1.8
Number of genes with >75% coverage
60
50
40
30
20
10
1
Number of genomes


C
0.56%
32.2%
56%
11.2%
Core
Softcore
Dispensable
Private


A
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
A
B
C
D
E
A. BTx623
B. Rio
C. RTx436
D. RTx430
E. Tx2783
B
CK
Tx2783
C
Number of variation/100kb
0
5
10
15
Chr01 Chr02 Chr03 Chr04 Chr05 Chr06 Chr07 Chr08 Chr09 Chr10
E
Color Key
Row Z-Score
SbiRTx2783.06G018300
SbiRTx2783.06G018800
SbiRTx2783.06G018900
SbiRTx2783.06G018400
SbiRTx2783.06G018500
Tx2783.5day.ck
Tx2783.5day.infested
Tx2783.10day.ck
Tx2783.10day.infested
Tx2783.15day.ck
Tx2783.15day.infested
D
SbiRTX2783.06G017600
SbiRTX2783.06G017700
SbiRTX2783.06G017800
SbiRTX2783.06G017900
SbiRTX2783.06G018000
SbiRTX2783.06G018200
SbiRTX2783.06G018300
SbiRTX2783.06G018400
SbiRTX2783.06G018450
SbiRTX2783.06G018460
SbiRTX2783.06G018500
SbiRTX2783.06G018600
SbiRTX2783.06G018700
SbiRTX2783.06G018800
SbiRTX2783.06G018900
SbiRTX2783.06G019000
SbiRTX2783.06G019100
SbiRTX2783.06G019200
SbiRTX2783.06G019300
SbiRTX2738.06G019400
Tx2783
SC326-6
SC237-14E
M35-1
Macia
PI656026
Malisor84-7
PI656056
R931945-2-2
ICSV745
PI561072
SC108C
PI276837
B963676
Cherekit
B923296
SC103-14E
Early Hegari
BTx623
SC56-14E
PI656065
PI656015
PI534047
RTx7000
PI656035
QL12
PI563516
Rio
PI329440
PI_653616
Karper 669
PI655978
SC170-6-8
IS8525
RTx436
RTx430
PI641824
SC62C
PI533998
PI653617
PI576399
IS9710
Ai4
LR9198
IBC/E-38432
PI585749
SC35C
SC23
B35
IS3614-2
PI655981
S.bicolor
Greenleaf
Zengada
PI330272
PI533902
PI525695
PI586430
PI300119
AusTRCF317961
S.propinquum 369-1
S.propinquum 369-2
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0


A
Color Key
Row Z-Score
−2
−1
0
1
2
Tx2783.5day.ck
Tx2783.5day.infested
Tx2783.10day.ck
Tx2783.10day.infested
Tx2783.15day.ck
Tx2783.15day.infested
B
Color Key
Row Z-Score
−2
−1
0
1
2
Tx2783.5day.ck
Tx2783.5day.infested
Tx2783.10day.ck
Tx2783.10day.infested
Tx2783.15day.ck
Tx2783.15day.infested
C
Cluster1
Cluster2
Cluster3
Cluster4
Cluster5
Cluster6
Cluster7
Cluster8
Cluster9
Cluster10
Cluster11
Cluster12
Cluster13
Cluster14
Cluster15
Cluster16
Expression change
5-day CK
5-day Infest
10-day CK
10-day Infest
15-day CK
15-day Infest