1 TITLE PAGE

1.1 TITLE

The *Sorghum bicolor* reference genome: improved assembly and annotations, a transcriptome atlas, and signatures of genome organization

1.2 AUTHORS

Ryan F. McCormick[1,2], Sandra K. Truong[1,2,], Avinash Sreedasyam[3], Jerry Jenkins[3], Shengqiang Shu[4], David Sims[3], Megan Kennedy[4], Mojgan Amirebrahimi[4], Brock Weers[2], Brian McKinley[2], Ashley Mattison[1,2], Daryl Morishige[2], Jane Grimwood[3,4], Jeremy Schmutz[3,4], and John Mullet[2]

1. Interdisciplinary Program in Genetics, Texas A&M University, College Station, TX 77843, USA
2. Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX 77843, USA
3. HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806, USA
4. Joint Genome Institute, Department of Energy, Walnut Creek, CA 94598, USA

1.3 CORRESPONDING AUTHOR

John Mullet: jmullet@tamu.edu

1.4 RUNNING TITLE

Sorghum bicolor version 3

1.5 KEYWORDS

genome assembly, reference genome, sorghum, nucleosome occupancy, gene annotation, Discrete Fourier Transform, genetic variation, satellite DNA, kinase

1.6 MANUSCRIPT TYPE

Resource

2 ABSTRACT

Sorghum bicolor is a drought tolerant C4 grass used for production of grain, forage, sugar, and lignocellulosic biomass and a genetic model for C4 grasses due to its relatively small genome (~800 Mbp), diploid genetics, diverse germplasm, and colinearity with other C4 grass genomes. In this study, deep sequencing, genetic linkage analysis, and transcriptome data were used to produce and annotate a high quality reference genome sequence. Reference genome sequence order was improved, 29.6 Mbp of additional sequence was incorporated, the number of genes annotated increased 24% to 34,211, average gene length and N50 increased, and error frequency was reduced 10-fold to 1 per 100 kbp. Sub-telomeric repeats with characteristics of Tandem Repeats In Miniature (TRIM) elements were identified at the termini of most chromosomes. Nucleosome occupancy predictions identified nucleosomes positioned immediately downstream of transcription start sites and at different densities across chromosomes. Alignment of the reference genome sequence to 56 resequenced genomes from diverse sorghum genotypes identified ~7.4M SNPs and 1.8M indels. Large scale variant features in euchromatin were identified with periodicities of ~25 kbp. An RNA transcriptome atlas of gene expression was constructed from 47 samples derived from growing and developed tissues of the major plant organs (roots, leaves, stems, panicles, seed) collected during the juvenile, vegetative and reproductive phases. Analysis of the transcriptome data indicated that tissue type and protein kinase expression had large influences on transcriptional profile clustering. The updated assembly, annotation, and transcriptome data represent a resource for C4 grass research and crop improvement.

3 INTRODUCTION

Sorghum bicolor, the fifth most important cereal crop in the world, is an economically important C4 grass grown for the production of grain, forage, sugar/syrup, brewing, and lignocellulosic biomass production for bioenergy. Meeting the food and fuel production challenges of the coming century will require production gains from traditional crop breeding, genomic selection, genome editing, and biotechnology approaches that develop plants with increased productivity and traits such as drought, pest and disease resistance, and canopies that have high photosynthetic efficiencies (Kromdijk et al., 2016; Mickelbart et al., 2015; Mondal et al., 2016; Mullet et al., 2014; Ort et al., 2015; Park et al., 2015; Technow et al., 2015; Voytas, 2013). Progress towards the genetic improvement of plants is promoted by the availability of foundational genetic and genomic resources. Because of this, we improved the *Sorghum bicolor* reference genome sequence assembly using targeted approaches and improved its annotation using data from a deep transcriptome analysis. A sorghum transcriptome atlas was created that contains gene expression data from the major plant tissue types across the juvenile, vegetative and reproductive stages of development. The genome sequence was used to analyze the distribution of key features in the genome including genes, transposable elements, genetic variation, and nucleosome occupancy likelihoods.

Sorghum is a diploid C4 grass with 10 chromosomes and an ~800 Mbp genome (Price et al., 2005). Cytogenetic and genetic analyses showed that sorghum chromosomes are comprised of distal regions of high gene density that exhibit high rates of recombination and large heterochromatic pericentromeric regions characterized by low gene density and low rates of recombination (Kim et al., 2005). A *Sorghum bicolor* reference genome sequence was reported in 2009, representing a major landmark in C4 grass genomics (Paterson et al., 2009). Reduced sequencing costs and technological advances have since enabled the sequencing and assembly of additional grass genomes, including *Brachypodium distachyon* (Vogel et al., 2010), corn (Schnable et al., 2009), foxtail millet (Bennetzen et al., 2012; Zhang et al., 2012), wheat (Brenchley et al., 2012), barley (Consortium, 2012b), and the desiccation tolerant *Oropetium thomaeum* (VanBuren et al., 2015). In addition, the genomes of 49 additional sorghum genotypes have been sequenced and assembled through alignment to the sorghum reference genome produced in 2009 (Evans et al., 2013; Mace et al., 2013; Zheng et al., 2011). Reference genomes provide an important resource for analyses, but their coverage and quality are often limited by the resources and technology available at the time of their construction. As such, reference genomes and their annotations benefit from iterative improvement as exemplified

by the Human genome project and related projects such as ENCODE (Consortium, 2012a; Consortium, 2004; Lander et al., 2001; Rosenbloom et al., 2013). To this end, we report an update to the BTx623 sorghum reference genome that leverages advances in sequencing technologies and transcriptomics to generate a more complete sorghum genome assembly and annotation.

A sorghum transcriptome atlas containing expression profiles of the major plant tissues was constructed to facilitate annotation of genes in the sorghum genome. Such atlas projects serve as resources for gene discovery, annotation, and functional characterization. Multiple atlas projects have been executed in recent years, including for maize and rice (Sekhon et al., 2013; Sekhon et al., 2011; Wang et al., 2010). In sorghum, microarray-based expression profiling and RNAseq have also been used to examine transcriptome dynamics in different sorghum genotypes, tissues, and responses to hormones and the environment (Abdel-Ghany et al., 2016; Shakoor et al., 2014). The current study contributes additional information on sorghum gene expression through construction of a sorghum transcriptome atlas using 47 samples collected from the major plant tissue types during the juvenile, vegetative and reproductive phases of plant development. Here we utilize the sorghum transcriptome atlas to facilitate gene annotation and to identify genes important for establishing organ identity in sorghum.

Additional features of the sorghum genome were investigated, including repetitive DNA elements, primary sequence-based nucleosome occupancy likelihoods, and the distribution of genetic variation among diverse sorghum accessions. Of particular interest was the identification of signatures that reflect higher-level organizational properties of the genome. Genetic variants do not accumulate uniformly across the genome due in part to regional variation in mutation rates (RViMR) that over time cause large differences in the number of genetic variants in different regions of eukaryotic genomes (Evans et al., 2013; Hodgkinson and Eyre-Walker, 2011; Makova and Hardison, 2015; Tolstorukov et al., 2011). In particular, chromatin structure has been associated with variation in the accumulation of genetic variants in human genomes (Tolstorukov et al., 2011). Additionally, previous work in medaka and humans found that genetic variation accumulated with a periodicity corresponding to nucleosome occupancy at transcription start sites (Higasa and Hayashi, 2006; Sasaki et al., 2009). Since nucleosome occupancy is associated with sequence identity, a support vector machine (SVM) was previously trained on human chromatin to predict nucleosome occupancy likelihoods from primary sequence, and the same SVM was shown to perform well in maize in predicting nucleosome occupancy (Fincher et al., 2013; Gupta et al., 2008). Given that eukaryotic

105 genomes are organized into higher order topologically associating domains and the influence of
106 nucleosome occupancy on the accumulation of genetic variation, the possibility that larger chromatin
107 domains influence the genome in a similar manner in plants also exists (Bonev and Cavalli, 2016).
108 As such, we explored the basis of genetic variation accumulation in the sorghum genome using
109 digital signal processing techniques.

110

111 # 4 RESULTS

112 ## 4.1 Genome assembly and improvement

113 Version 1 of the sorghum BTx623 reference genome assembly incorporated 625.6 Mbp of genomic
114 sequence into 10 pseudomolecules corresponding to the 10 sorghum chromosomes by combining
115 data from whole genome shotgun sequencing and targeted sequencing of BACs and fosmids using
116 paired-end Sanger sequencing,. An error rate of < 1 per 10 kbp was estimated based on Sanger
117 sequencing of BACs (Paterson et al., 2009). Version 2 of the sorghum reference genome assembly
118 was publicly released without a corresponding publication; as such, all comparisons here are made
119 relative to version 1.

120 In this study, version 1 of the sorghum reference genome was refined by deep whole genome short
121 read sequencing (110X) and targeted finishing of gene-dense regions of the genome (greater than 2
122 genes per 100 kbp) using primer walking via Sanger sequencing and shotgun sequencing of plasmid
123 subclones, fosmid, and BAC clones (Supplemental File S1). These finished regions were assembled
124 and hand-curated (representing 344.4 Mbp), mapped back to the v1 assembly, and then incorporated
125 into the v1 assembly, adding a total of 4.96 Mbp to the assembly. To improve ordering of the
126 reference genome, a high-density genetic map based on ~10,000 markers genotyped in a 437-line
127 recombinant inbred mapping population derived from the sorghum lines BTx623 and IS3620C was
128 used to integrate 7 additional scaffolds into chromosomes (Truong et al., 2014). Furthermore, the
129 genetic map identified a 1.08 Mbp region that was previously assembled into chromosome 6, but
130 markers within the region were not linked to flanking regions on chromosome 6 and tightly linked
131 with markers on chromosome 7 (Figure 1). This assembly error in version 1 is corrected in version 3.

132

133

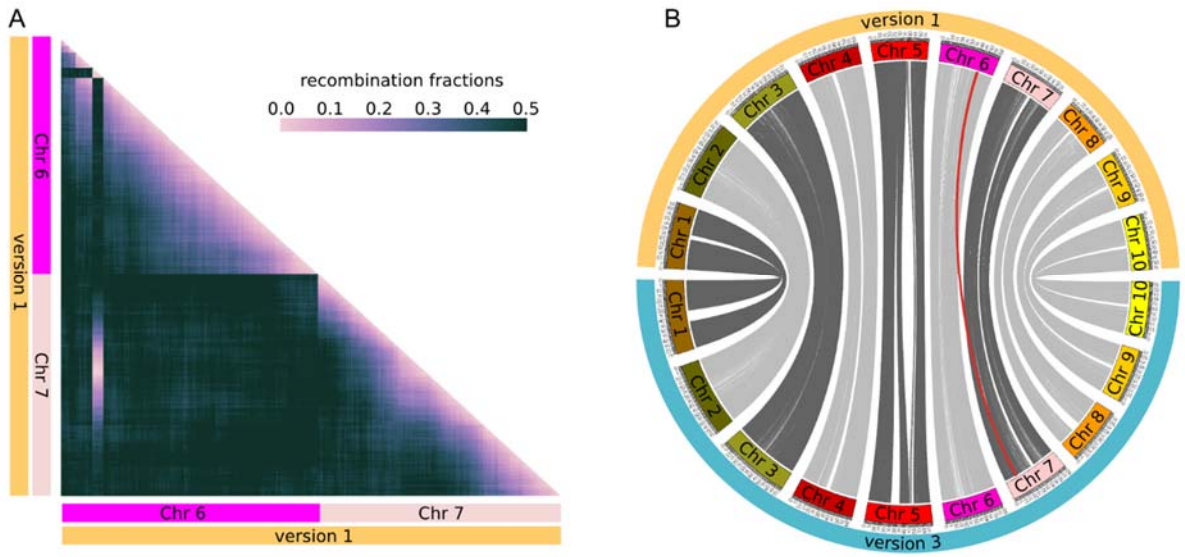

Figure 1: Correction of misassembled region in the version 1 sorghum reference genome assembly and integration of new sequence. (A) Recombination fractions of markers in the BTx623 x IS3620C sorghum recombinant inbred line (RIL) population ordered by physical position relative to the version 1 reference assembly. A block of markers spanning roughly 1 Mbp were previously physically assembled on chromosome 6, but are genetically unlinked with markers on chromosome 6. Instead, the markers are tightly linked with a region of chromosome 7. (B) Sequence identity mapped between the version 1 and version 3 of the reference assemblies. A 1.08 Mbp region previously located on chromosome 6, corresponding to the markers in panel A, was moved to chromosome 7. Additional sequences were integrated into the chromosomes, expanding the size of the version 3 assembly (Supplemental File S1).

Due to integration of additional sequence during finishing and of previously unplaced contigs into the main genome sequence, the contiguity of the v3 sequence comprising the 10 sorghum chromosomes increased significantly, such that the N50 length, the largest length such that 50% of all bases are contained in contigs of at least that length (Lander et al., 2001), increased by 6.3 fold from 0.2045 Mbp to 1.5 Mbp. The resulting v3 assembly included 655.2 Mbp of genomic sequence incorporated into chromosomes, with an estimated error rate of <1 per 100 kbp (Table 1).

Table 1: Summary statistics for sequence comprising the 10 chromosomes for the version 1 and version 3 reference assemblies. The number of bases incorporated into the genome, the contiguity of the sequence, and the accuracy of the sequence improved in version 3. N50 is defined as the largest length such that 50% of all bases are contained in contigs of at least that length (Lander et al., 2001), and L50 is defined as the number of contigs, where, when summed longest to shortest, the sum exceeds 50% of the assembly size.

	Sorghum bicolor reference genome pseudomolecules	
	Version 1	Version 3
Number of pseudomolecules	10	10
Number of contigs	6,929	2,688
Scaffold sequence (Mbp)	659.2	683.6
Contig sequence (Mbp)	625.6	655.2
Scaffold N50 (Mbp)	64.3	68.7
Contig N50 (Mbp)	0.2045	1.5
Scaffold L50	5	5
Contig L50	838	71
Unmapped sequence (Mbp)	71.9	20.2
Estimated error rate	< 1 per 10 kbp	< 1 per 100 kbp

4.2 Annotation of genes and other features in the sorghum genome.

The version 3 (v3.1) assembly was annotated for a number of feature types, including genes, repetitive elements, genetic variation, and primary sequence-based nucleosome occupancy predictions (Figure 2, Supplemental Figures S1 and S2). Deep transcriptome profiles were obtained from 47 different tissues or developmental phases to facilitate the annotation of genes in the sorghum genome. Tissues from growing and developed portions of roots, leaves, stems, seeds, and panicles were isolated during the juvenile, vegetative, and reproductive phases of plant development. Illumina sequencing of cDNA obtained from these tissue samples (RNA-seq) generated 3.3 billion sorghum paired-end reads. The sequence reads were subsequently combined with sorghum ESTs and homology-based predictions to annotate 34,211 genes in the *Sorghum bicolor* genome (gene set version 3.1). The v3.1 gene annotation represents a 24% increase relative to the 27,607 genes annotated in version 1 (gene set version 1.4). The median and mean gene size in v3.1 increased to 1600 and 1835, from 1336 and 1473 in v1.4, respectively, due primarily to improved annotation of

exons. As such, the number of genes, as well as the length of genes increased significantly indicating that the v3.1 gene annotation is the most comprehensive sorghum gene annotation to date. A small number (175) of genes in v1.4 were not supported and were not included in the v3.1 gene set. Repetitive elements in the sorghum genome were annotated using a *de novo* repetitive element annotation pipeline in conjunction with existing repetitive element libraries (Bao et al., 2015; Flutre et al., 2011; Ouyang and Buell, 2004; Quesneville et al., 2005). Consistent with the previous annotation of the v1 assembly, the percentage of the genome annotated as retrotransposons (i.e. class I elements) was 58.8%, most of which were long terminal repeats (54% of the genome). Approximately 8.7% of the genome annotated as DNA transposons (i.e. class II elements).

The distributions of genes, repetitive elements, and genetic variants across each sorghum chromosome were generated using 1Mbp sliding windows (Figure 2, Supplemental Figures S1 and S2). Genes are at higher density in the distal euchromatic regions of chromosome arms and repetitive sequences related to transposable elements are most dense in heterochromatic pericentromeric regions characteristic of sorghum chromosomes (Evans et al., 2013; Paterson et al., 2009). The accumulation of genetic variation in *Sorghum bicolor* accessions was examined by aligning and comparing reads from 56 resequenced sorghum genotypes to the v3 genome sequence. *Sorghum propinquum* samples and two subsp. *verticilliflorum* genotypes were removed before analyses of variant distribution due to their evolutionary divergence from BTx623 and other resequenced *Sorghum bicolor* genotypes. The analysis identified 7,375,006 single nucleotide polymorphisms (SNPs) and 1,876,974 insertion/deletions (indels) distributed across the 10 chromosomes. The density of genetic variants was highly variable across the sorghum genome, with higher variant density in the distal euchromatic regions relative to heterochromatic pericentromeric regions of each chromosome, consistent with previous reports (Evans et al., 2013).

Predicted nucleosome positioning in the BTx623 v3 reference genome was examined by generating nucleosome occupancy likelihoods using a support vector machine trained on human chromatin data and validated in maize. Using this approach every nucleotide position was assigned a nucleosome occupancy likelihood (NOL) based on the primary sequence identity of a 50 bp window centered on the nucleotide (Fincher et al., 2013; Gupta et al., 2008). While primary sequence is not the only determinant of nucleosome binding, it influences the relative affinity of binding and general trends are indicative of chromatin organization. The predicted nucleosome occupancy likelihoods for sorghum are similar to maize in that the distributions vary across each chromosome, but with a

relatively uniform pattern that does not match variation in gene or repeat density across each chromosome (Figure 2, Supplemental Figures S1 and S2).

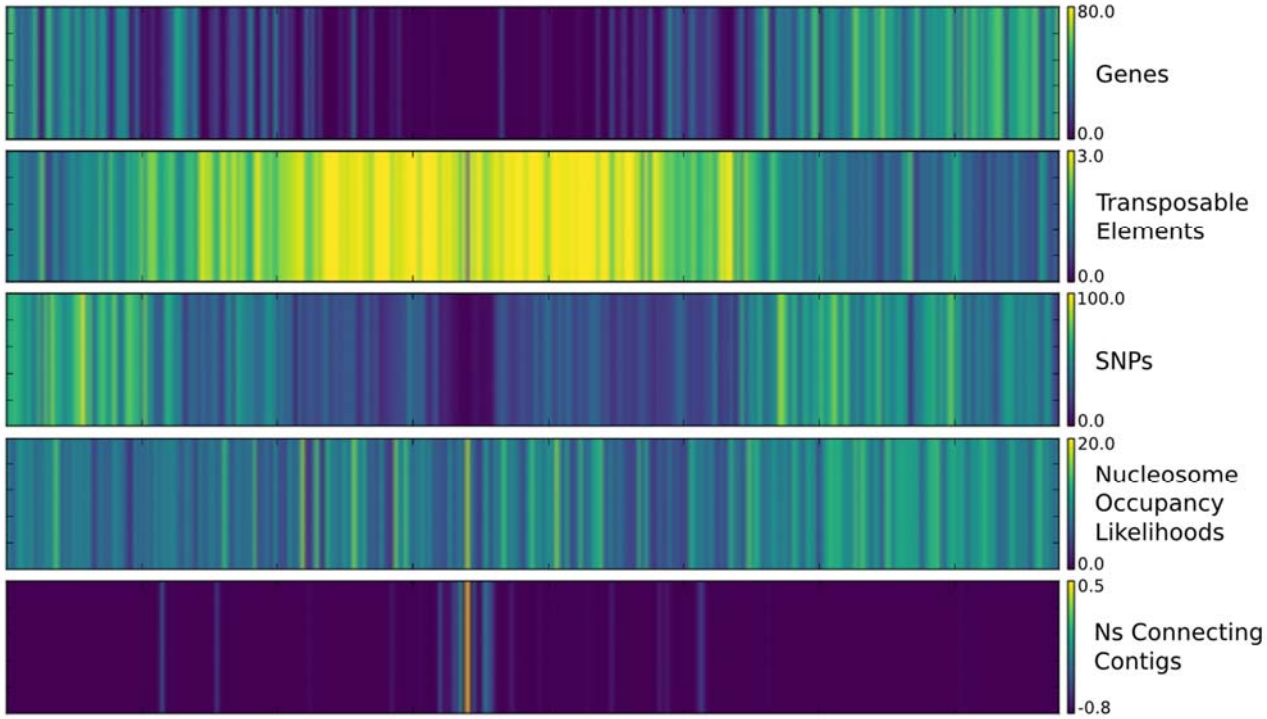

Figure 2: Feature densities and score averages across chromosome 2 of the sorghum genome. Color map displaying the average densities of multiple features across chromosome 2 of the sorghum genome, including annotated genes, transposable elements, single nucleotide polymorphisms, nucleosome occupancy likelihoods, and uncalled bases (Ns) connecting contigs in the assembly. Maps for all 10 chromosomes are depicted in Supplemental Figures S1 and S2.

4.3 Periodicity in features related to variant distributions in the sorghum genome

Information in eukaryotic genomes is stored at multiple scales, ranging from single base-pairs that specify codon identity to megabase-sized topologically associated domains that regulate transcriptional states (Bonev and Cavalli, 2016). Some of these organizational properties are correlated with periodic signatures in the accumulation of genetic variation. For example nucleosome positioning generates periodicity in the accumulation of genetic variants in humans and medaka (Higasa and Hayashi, 2006; Sasaki et al., 2009; Tolstorukov et al., 2011). Given that these organizational properties are associated with genomic signals such as variant density, digital signal

processing techniques can be used to identify signatures associated with these properties. To this end, the Discrete Fourier Transform (DTF) was used to examine periodicities in the accumulation of genetic variation and nucleosome occupancy likelihoods to help identify mechanisms by which the sorghum genome stores information.

A known functional feature of the genome that influences the accumulation of genetic variation is the wobble base in codons. Due to redundancy in the genetic code, every third base downstream of a coding sequence start site is under relaxed selection since the primary DNA sequence is often able to change without dramatically influencing the information content of the sequence. This manifests as a prominent periodicity with a period of 3 bp after processing the polymorphism accumulation signal in the coding sequence of sorghum genes for regions downstream of coding start sites, but not upstream (Figure 3).

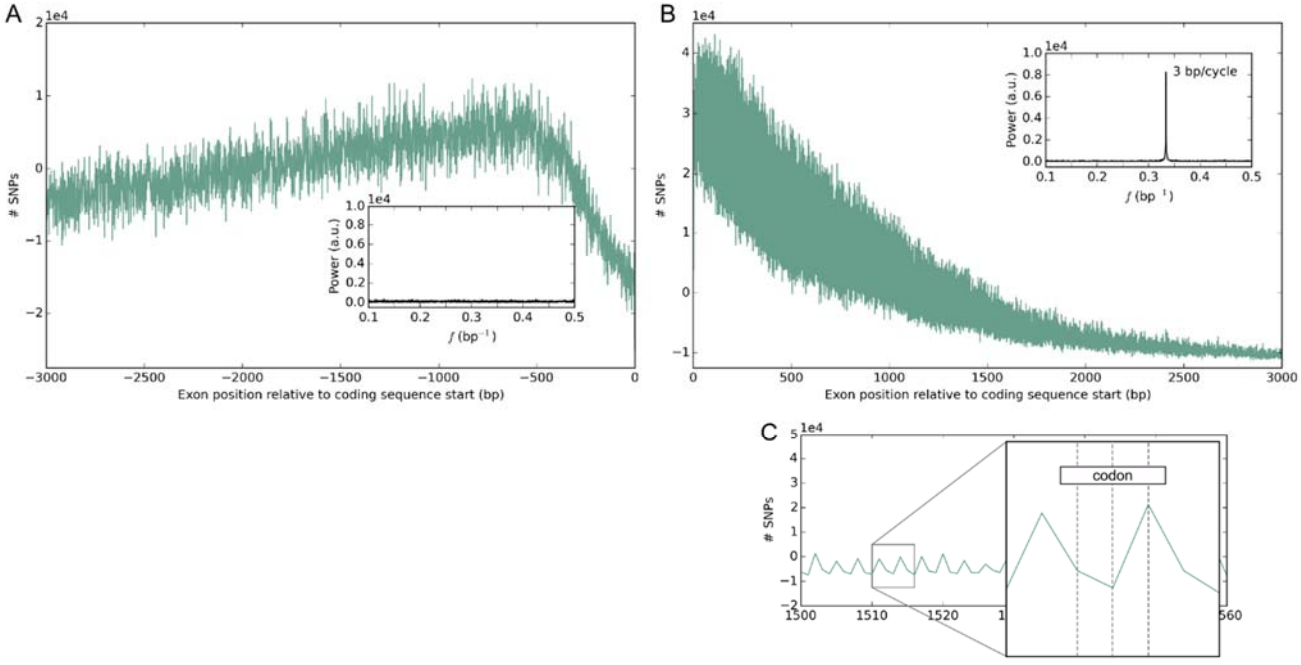

Figure 3: Functional properties of the sorghum genome leave periodic signatures that can be identified using signal processing techniques. Due to the degeneracy of the genetic code, relaxed selection at the wobble base in codons causes SNPs to accumulate with a periodicity of 3 bp downstream of coding sequence start sites in exon sequence (B), but not upstream of coding sequence start sites in the sorghum genome (A). This manifests as a strong signal at 0.33 bp^{-1} after transforming the SNP accumulation signal with the DFT (inset of B). (C) Zoom in of panel B shows the periodic signal. The Y axis of panels A and B plot the number of SNPs relative to the average of the respective window. The Y axis represents the sum of SNPs at each position relative to the CDS start site across all genes in the genome, centered to

the mean of the respective window; CDS lengths of less than 3000 were considered to have 0 SNPs between their end and 3000 bp, leading to the apparent decline observed in panel B.

Nucleosome scale variant periodicities were examined for signatures of genome organization because studies in medaka and human indicated that genetic variation accumulates at transcription start sites (TSSs) with periodicities around 150 bp, corresponding to nucleosome occupancy (Higasa and Hayashi, 2006; Sasaki et al., 2009). To determine if a similar phenomenon was present in the sorghum genome, the genetic variation that accumulated around transcription start sites as well as nucleosome occupancy likelihoods were examined. Consistent with micrococcal nuclease digestion results in maize and *Arabidopsis,* prediction scores indicated a high likelihood of a nucleosome positioned immediately downstream of the transcription start site of genes in sorghum (Figure 4) (Fincher et al., 2013; Liu et al., 2015). While variant frequency decreased immediately downstream of TSSs, the variant profile in sorghum did not show accumulation of genetic variants with a period of ~150 bp downstream of these sites. Nucleosome occupancy predictions also did not predict a periodic arrangement of nucleosomes downstream of transcription start sites.

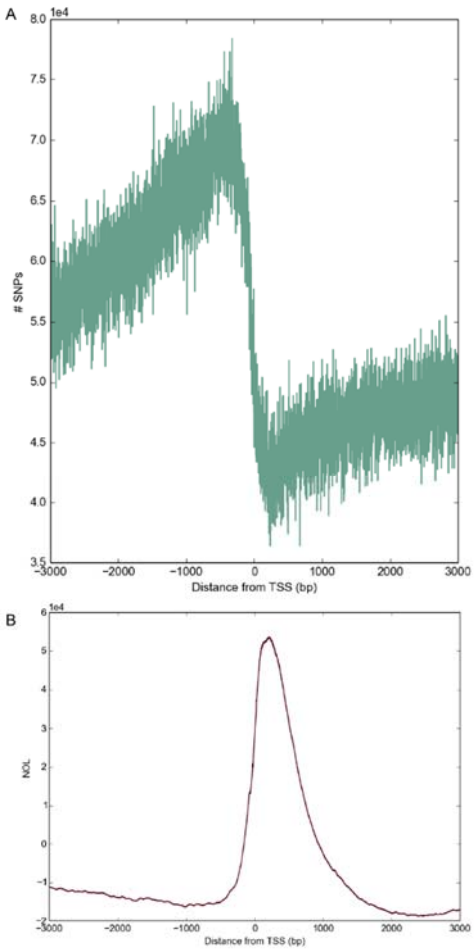

Figure 4: Genetic variation and nucleosome occupancy likelihoods around transcription start sites in the sorghum genome. Nucleosome occupancy scores indicate a high likelihood of a nucleosome positioned immediately downstream of transcription start sites in sorghum. Strong evidence that nucleosomes were stably positioned based and periodically arrayed as in medaka and human was not observed in either the accumulation of genetic variants nor nucleosome occupancy likelihoods, though NOLs indicate that a nucleosome is often positioned immediately downstream of the transcription start site, consistent with experimental observations in maize and *Arabidopsis*.

Nucleosome scale periods of 180 bp are present in nucleosome occupancy likelihood profiles in multiple regions of the genome, and are especially pronounced in subtelomeric regions, suggesting the possibility of stably positioned, periodically arrayed nucleosomes downstream of the $(CCCTAAA)_n$ telomere repeats present at the end of sorghum chromosomes (Figure 5B and 5C) (Klein et al., 2000).

Since the SVM used for nucleosome occupancy likelihood calculation used only primary sequence, any primary sequence that was tandemly arrayed (e.g., satellite DNA) should also yield a periodic

signal. Further characterization of the primary sequence underlying the periodic signal identified that the periodicity indeed resulted from tandemly arrayed, subtelomeric, satellite DNA with a repeat size of 180 bp, consistent with observations that the monomer size of satellite DNA repeats often correspond to the length of DNA wrapped around nucleosomes (Mehrotra and Goyal, 2014). BLAST analyses indicated that most chromosome arms contained tandem arrays of one of two satellite repeats, with the two types of repeats sharing some sequence identity (Figure 5A). The two monomers are referred to as subtelomeric tandemly arrayed 1 and 2 (STA1 and STA2) here for brevity.

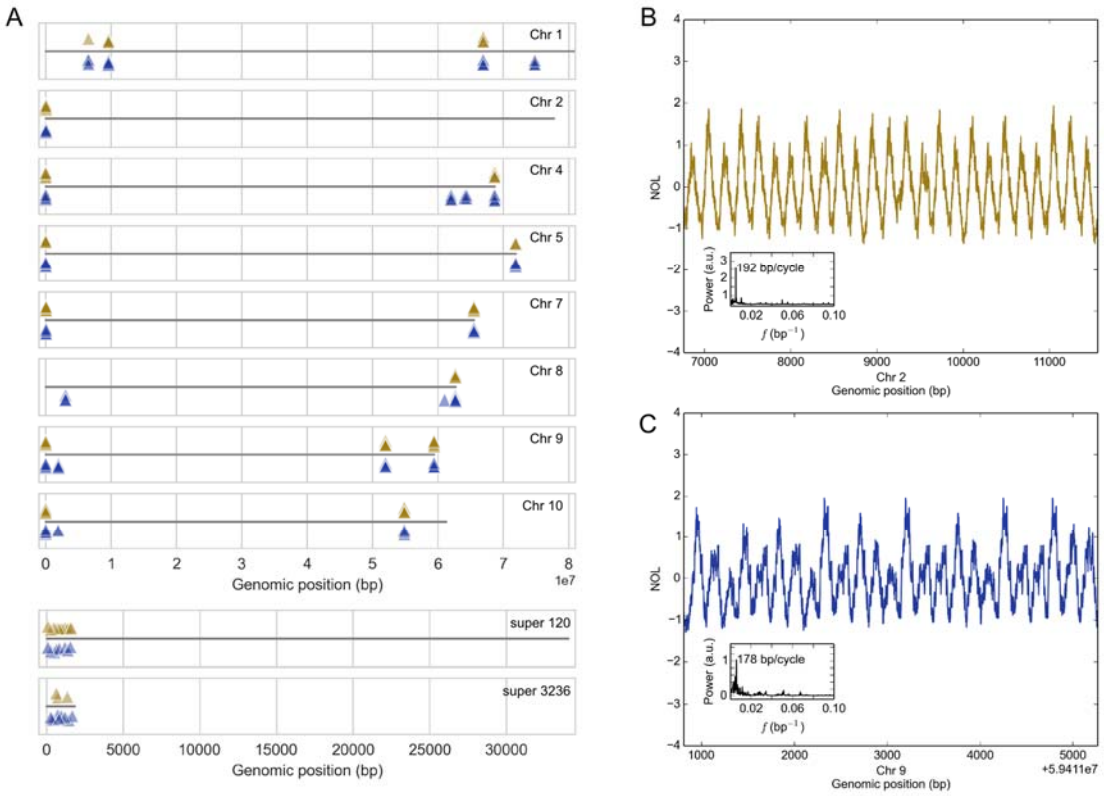

Figure 5: Subtelomeric periodicities in nucleosome occupancy likelihoods correspond to arrays of tandem repeats located near the end of most chromosome arms. (A) Graphic representation of BLAST hits for the consensus sequence of STA1 and STA2 indicate that most chromosome arms contain subtelomeric tandem arrays of the STA1 or STA2 monomer; two super contigs in the assembly also contain arrays, and may correspond to subtelomeric sequence on the arm of chromosome 2. (B) Nucleosome occupancy likelihoods (centered on the mean) and power spectrum for an array of the STA1 monomer with multiple sequence alignment of continuous arrays from multiple chromosome arms. (C) Same as panel A, but with arrays of the STA2 monomer. STA1 and STA2 share sequence identity and are likely related, though most chromosome arms bear tandem arrays of only one or the other; BLAST hits show colocalization due to shared identity.

Tandem arrays of STA1 or STA2 (or a complex mixture of both) exist on most of the sorghum chromosome arms, with the longest array present at the beginning of chromosome 2, repeating STA1 more than 200 times over more than 36 kbp. Arrays of STA1 or STA2 are present within 50 kbp of the beginning and end of chromosomes 4, 5, 7, and 9. Chromosomes 3 and 6 are the only scaffolds without the elements near the ends of one of the chromosome arms (Figure 5). Notably, the arrays are also found on super contigs 120 and 3236; these may correspond to the ends of one or more chromosomes, although they lack the $(CCCTAAA)_n$ telomeric repeat. Telomeric repeats were found at both termini of chromosomes 1, 4, 5, 7 and 10 and at one of the two termini of chromosomes 2, 3, 6, 8 and 9, so no strong relationship between the presence of an assembled telomere and the STA repeat was observed (Supplemental Table S1).

Alignment searches for STA1 and STA2 in maize, rice and more distantly related plants suggest that this sequence repeat feature is sorghum specific. *De novo* repetitive element annotation identified the arrays as individual terminal-repeat retrotransposons in miniature (TRIM) elements, although they were not included in a recent annotation of plant TRIMs, a database that includes sorghum (Gao et al., 2016). While TRIMs have been observed to accumulate in tandem arrays, the monomers of STA1 and STA2 lack most of the features of canonical TRIM elements (Gao et al., 2016; Witte et al., 2001). Only STA1 bears a putative primer binding site (PBS; complementary to the sorghum methionine tRNA). Notably, STA1 shares sequence identity with an unclassified sorghum element (SRSiOTOT00000007) from the TIGR Plant Repeat Database (Ouyang and Buell, 2004), as well as the *S. halepense*-specific repetitive elements XSR6, XSR1, and XSR3 (Hoang-Tang et al., 1991). The STA1 and STA2 monomers both have a complex substructure of internal duplication and tandem repeats (Figure 6, Supplemental Figure S3, and Supplemental File S2).

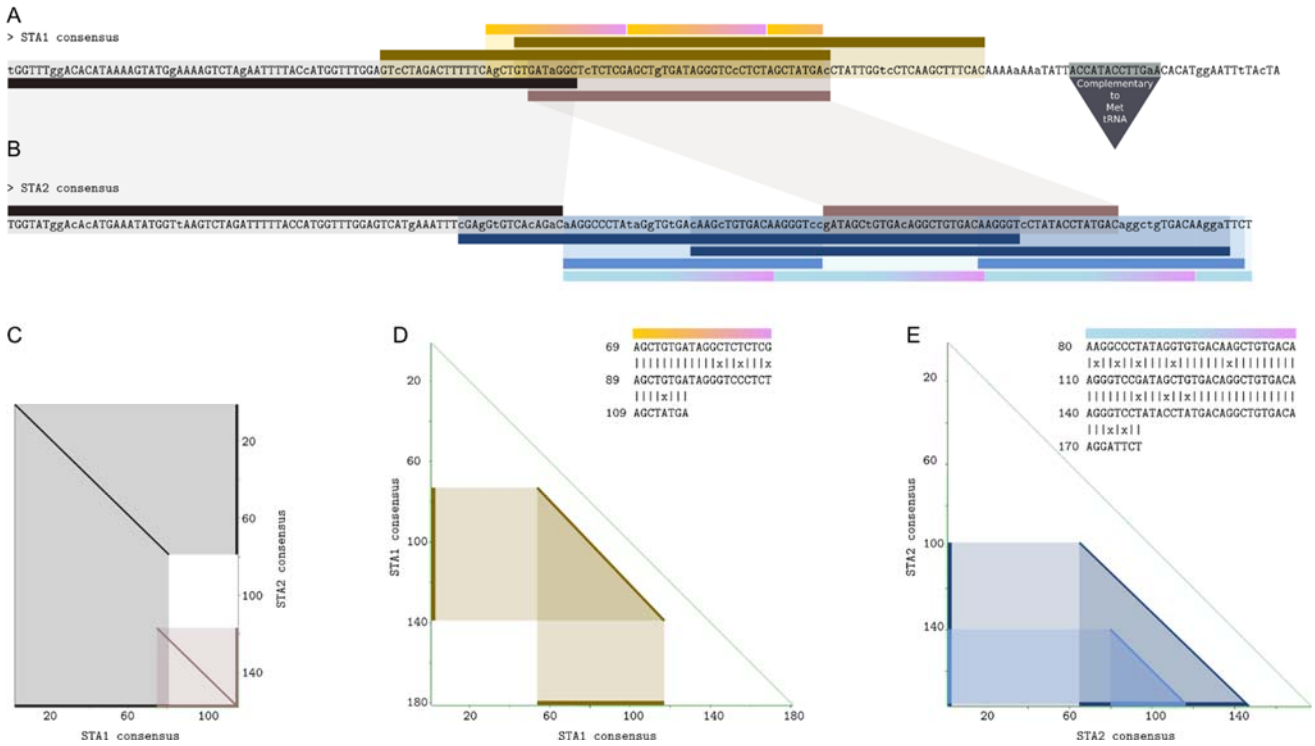

Figure 6: Relationships and structure of STA monomers. Monomers of STA1 and STA2 share blocks of identity and each contain direct repeats. (A, B) Consensus sequence of the STA1 and STA2 monomers annotated for sequence shared between monomers as well as direct repeats within each monomer. Annotations are indicated by colored bars and color coded according to their origin, either from (C) an alignment dotplot of STA1 and STA2, (D) a self-alignment dotplot and tandem direct repeat analysis of STA1, or (E) a self-alignment dotplot and tandem direct repeat analysis of STA2. STA1 also contains a putative promoter binding site (PBS) complementary to a sorghum methionine tRNA, a feature characteristic of TRIM elements (Witte et al., 2001).

Signatures generated as a consequence of genome regulation or structure can be detected by signal processing techniques as shown previously for the effects of the wobble position in codons (Figure 3). Therefore, variant accumulation at larger scales was examined across the genome to identify periodic patterns of SNP accumulation. Consistent with previous reports, DNA variant profiles obtained from comparison of sorghum genome sequences from genotypes representing different sorghum races revealed that variants accumulate in a non-uniform fashion across the genome (Evans et al., 2013). Genome-wide analysis of SNP accumulation using the DFT indicated that this non-uniformity occasionally manifested as a periodic event where genetic variants accumulated in peaks and troughs in a region of the genome. Genome-wide scans of variant accumulation using the DFT indicated that multiple regions of the genome display large-scale periods, such that a peak of variant

accumulation is observed every 25 kbp (Figure 7). As with the periodicity observed at the wobble base, the cyclical nature of peaks in variant accumulation may represent a consequence of genome organization or information storage. This large scale periodicity of SNP accumulation was observed in regions of chromosomes 1, 3, 4, 5, 9, and 10 when SNPs called from sequence data for 52 sorghum genotypes were analyzed.

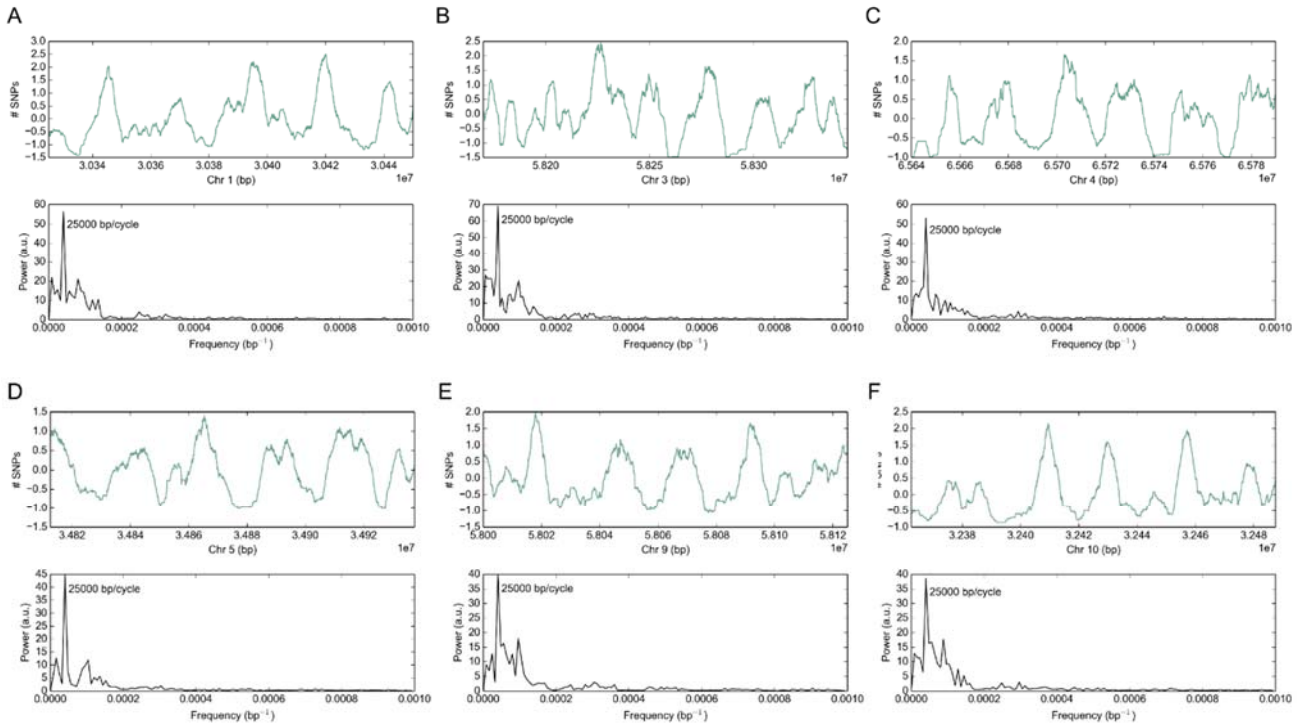

Figure 7: Periodicities in the accumulation of genetic variation in the sorghum genome. A genome-wide scan for periodic accumulation of SNPs identified multiple regions of the genome with a distinct period of 25,000 bp. The top plot of each panel shows the accumulation of SNPs relative to the mean of the window given a 5,000 bp sliding average, and the bottom plot shows the power spectrum after transformation with the Discrete Fourier Transform (A-F).

4.4 The Sorghum Transcriptome Atlas

The sorghum transcriptome atlas used to improve the sorghum reference genome gene annotation represents a broad diversity of tissues, developmental stages, and responses to nitrogen sources, encompassing a variety of transcriptional states. The transcriptome atlas was developed with two primary goals: (1) to sample the major plant organs (roots, leaves, stems, panicles) when these organs

were growing and then following maturation at different developmental stages (juvenile, vegetative, reproductive) to facilitate comprehensive annotation of genes in the sorghum genome and (2) to sample a diversity of nitrogen states and sources as part of an inter-species plant gene atlas project. A thorough analysis of these datasets is beyond the scope of this manuscript, but they are described here for release into the public domain for use by the community at large. The samples collected are described in Supplemental Table S2 and Supplemental File S3.

Initial analyses of the transcriptome data were carried out to provide a high-level overview of the transcriptome atlas contents. Correlations of the expression values across all 34,211 genes indicated high correlation within biological replicates of the same sample, as well as correlated groups between samples from the same tissue (Supplemental Figure S4). The largest block of correlated expression was a block of high correlation between all of the root samples, regardless of whether the root sample was more distal or proximal or root nitrogen treatment. Dormant seed shared the least correlation with any of the samples, indicating that its steady state pool of transcripts differed the most dramatically from other tissues analyzed.

Hierarchical clustering based on the transcript abundance of all 34,211 genes via UPGMA identified similar relationships among the samples, indicating that the transcript pool of a given sample was defined predominantly by the tissue/organ identity rather than the developmental stage. Seed samples were the most transcriptionally distinct, especially dormant seed. In agreement with hierarchical clustering, k-means clustering indicated that roots, stems, leaves, and seeds formed distinct clusters based on gene expression (Figure 8).

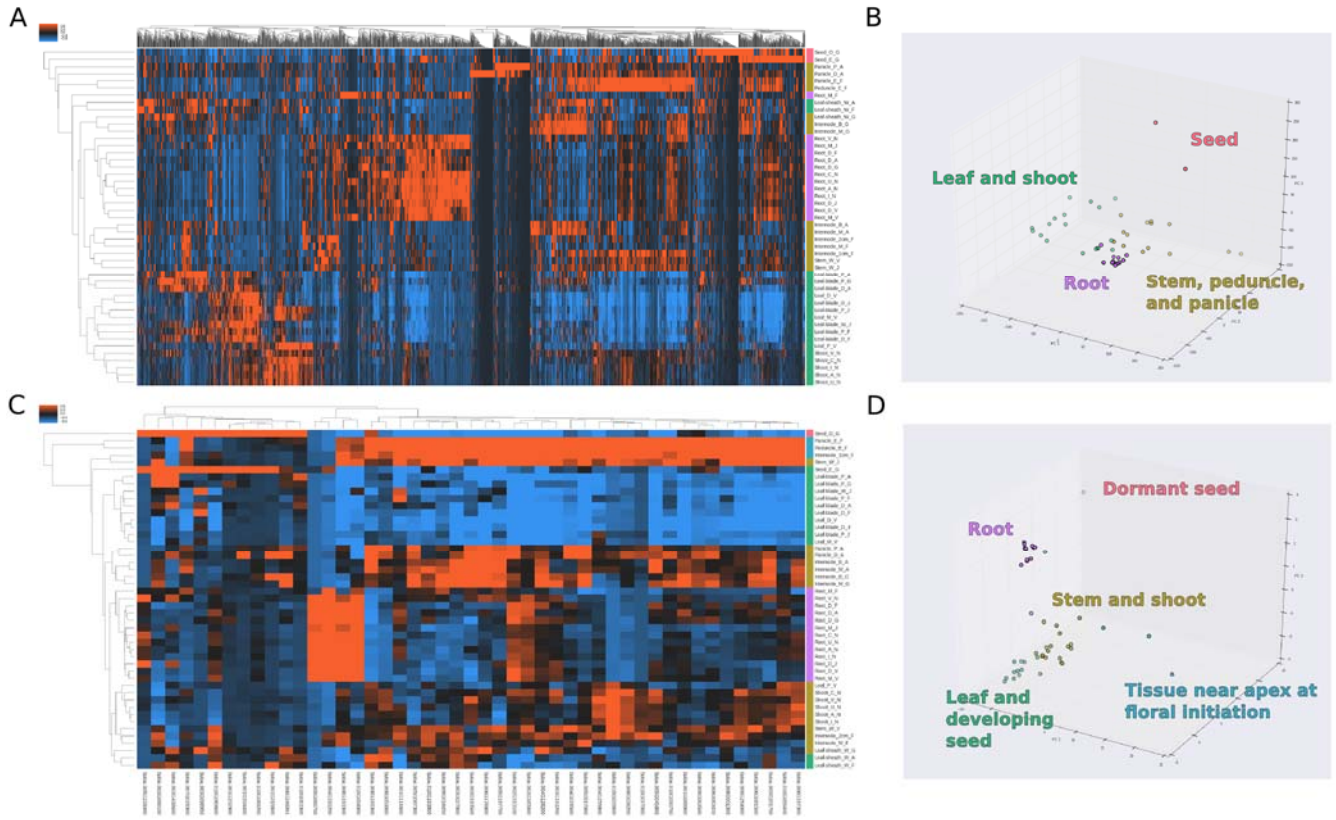

Figure 8: Clustering and ontological analyses indicate the expression of kinase genes are associated with tissue identity. (A) Heat map and hierarchical clustering of atlas samples based on gene expression of all 34,211 sorghum genes; color bars on right correspond to k-means clusters in panel B. (B) Scores of the first three principal components of the atlas samples colored based on k-means cluster (k = 4) using expression values of all genes. (C) Ontological enrichment analysis of the 2,500 genes with the largest loadings for the first three principal components indicate that kinase genes were overrepresented, and the expression of the 47 kinase genes driving the enrichment are plotted as a heat map with hierarchical clustering; color bars on the right correspond to k-means clusters in panel D. (D) Scores for the first three principal components of the atlas samples colored based on k-means cluster (k = 5) using expression values of the 47 kinase genes.

To identify a set of genes with large variation in expression across the dataset, principal component analysis was performed using all 34,211 genes to obtain the first three principal components (PCs), and the set of 2,500 genes with the largest sum magnitude of loadings for the first three PCs were identified (Supplemental Figure S5). To determine if particular classes of genes were overrepresented among these genes that explained large components of variation in the dataset, ontological enrichment analysis was performed for terms related to molecular functions. Three molecular function enrichments were identified, including structural constituents of the ribosome (GO0003735),

388 protein kinase activity (GO0004672), DNA-directed RNA polymerase activity (GO0003899)
389 (Supplemental File S4). Since kinase activity is associated with signal transduction and the other two
390 terms may be explained by the developmental state of the tissue (e.g., high transcription and
391 translation activity), the 47 genes responsible for protein kinase activity were investigated further
392 (Supplemental Table S3).

393 Hierarchical clustering and k-means clustering based on the expression state of the 47 kinase activity
394 genes broadly reproduced the same groupings as all 34,211 genes (Figure 8). Notably, three samples
395 associated with proximity to the shoot apical meristem at floral initiation were clustered together,
396 potentially representing kinases involved in the transition from a shoot apical meristem to a floral
397 meristem.

398

399 **5 DISCUSSION**

400 The sorghum genome sequencing project was organized in 2005 because *Sorghum bicolor* has a
401 relatively simple genome compared to many other grasses, sorghum is a valuable genetic model for
402 C4 grass research, and sorghum crops are important world wide, especially as subsistence crops in
403 the semi-arid tropics (Participants, 2005). The sorghum genome sequence improved our
404 understanding of sorghum genome organization, coding capacity, and aided analysis of grass genome
405 diversification (Paterson et al., 2009). The original reference genome sequence was based on ~8.5-
406 fold depth paired-end Sanger sequence reads from genomic libraries with 100-fold variation in the
407 size of inserts. Discrepancies in order that arose during assembly were resolved in part using
408 information from pre-existing high resolution genetic and BAC-based physical maps of the sorghum
409 genome (Bowers et al., 2003; Klein et al., 2000). The sum of the 201 largest sequenced scaffolds
410 spanned 678.9 Mbp of which 625.7 Mbp of the sequence was assigned chromosomal locations. The
411 size of the reference genome had been previously estimated by flow cytometry to be 818 Mbp (Price
412 et al., 2005), indicating that reference genome v1 sequence comprising the 10 chromosomes
413 accounted for ~76% of the total genome sequence. It was reported that 15 of the 20 chromosome
414 termini contained telomeric repeats and that Cen38 sequences (Zwick et al., 2000) were present in
415 each chromosome, although these sequences were also found in many of the sequence scaffolds that
416 could not be incorporated into the chromosomal sequences (Paterson et al., 2009). Despite the need
417 for further improvement, the resulting sorghum reference sequence has been of great value to the

418 sorghum and grass research community, enabling comparative genomics (Paterson et al., 2009),
419 association studies (e.g., Brenton et al., 2016; Morris et al., 2013), the development of genotyping by
420 sequencing methods for sorghum (Morishige et al., 2013), analysis of sorghum diversity and variant
421 distribution (Evans et al., 2013; Mace et al., 2013; McCormick et al., 2015), genome methylation
422 profiles (Olson et al., 2014), and many other research activities.

423 The objective of the current study was to update the sorghum reference genome sequence and its
424 annotation, and to characterize additional features of the sorghum genome that affect sorghum
425 biology. The sequence quality and coverage of the reference genome was improved by obtaining
426 110X coverage of the genome using Illumina sequencing, targeted finishing of ~344 Mbp of gene
427 rich portions of the genome, and by improving order and sequence contiguity using a high density
428 genetic map. These activities increased sequence coverage by ~30 Mbp, reduced error frequency 10-
429 fold to ~1/100 kbp, and improved assembly order by moving a 1 Mbp block of DNA from SBI-06 to
430 SBI-07. The research did not identify and incorporate sequences containing telomeric repeats that
431 are missing from the ends of 5 chromosomes and the order and completeness of sequences in the
432 pericentromeric regions that have high repeat density was not significantly changed. Long read
433 sequencing and Hi-C analysis (Sanborn et al., 2015) would be valuable approaches to implement to
434 further improve the reference genome sequence.

435 Version 1.4 of the sorghum genome sequence provided evidence for 27,604 annotated genes.
436 Subsequent analysis of gene annotations that incorporated RNA-seq data indicated that a large
437 number of genes were not annotated in v1.4 and that many of the annotations were incomplete
438 (Olson et al., 2014). Results from the current study based on deep RNA-seq analysis of 47 tissues
439 from roots, stems, leaves, leaf sheaths, panicles and seed enabled the annotation of 34,211 genes, a
440 24% increase relative to v1.4. RNA-seq data also improved the annotation of exons resulting in a
441 significant increase in average gene size consistent with prior results based on a similar approach
442 (Olson et al., 2014). Increased gene coverage and improved gene annotation and sequence accuracy
443 will aid comparative genomics studies as well as GWAS and map-based QTL to gene discovery
444 projects that can result in false negatives/positives if the reference genome sequence used for analysis
445 is not a well annotated high quality sequence. In our own research, errors and misannotation of the
446 v1 sequence caused identification candidate gene alleles underlying QTL to be missed until direct
447 sequencing was carried out on all genes in fine mapped intervals (Hilley et al., 2016; Murphy et al.,
448 2011).

While v3.1 is a substantial improvement over v1, additional information is needed to fill in missing portions of the genome sequence and to improve gene annotation. As noted above, one end of 5 chromosomes lack telomeric sequences indicating these chromosome sequences are not complete. Moreover, it is likely that the sequence of the pericentromeric repeat-rich regions of chromosomes is incomplete and possibly misordered in some regions. Since recombination is extremely low across the large heterochromatic pericentromeric regions (Kim et al., 2005), the high resolution genetic map employed to order DNA in euchromatic regions was not useful for ordering sequences across the pericentromeric regions. A combination of long range, long read sequencing and Hi-C analysis would be useful to improve these regions of the reference genome. In addition, Iso-Seq was shown to aid the analysis of full-length splice isoforms, alternative polyadenylation sites, and non-coding RNAs in sorghum (Abdel-Ghany et al., 2016). The analysis showed that in depth Iso-Seq data will significantly improve the current annotation of the sorghum genome and transcriptome. Moreover, pan-genome projects in maize and other species show that a substantial number of 'dispensible' genes are found only in a subset of the genotypes of a species germplasm (Hirsch et al., 2014). Therefore characterization of the sorghum pan-genome will require the acquisition and de novo assembly of genomes from diverse sorghum genotypes possibly aided by the construction of a set of reference genomes sequences that sample sorghum's diversity space.

The distribution of genes, repeats, variants, and other features of the sorghum genome was updated based on the v3 genome sequence. Gene density was highest in distal euchromatin portions of chromosomes and repetitive sequences related to retrotransposons were enriched in heterochromatic pericentromeric regions as previously described (Kim et al., 2005; Paterson et al., 2009). Predicted nucleosome positioning based on primary sequence data showed localized variation in nucleosome density but a fairly uniform distribution of nucleosome localization across chromosomes. Digital signal processing of genomic signals is a useful approach to identify novel patterns in genome structure. Through this approach, previously uncharacterized subtelomeric tandem repeats were identified in sorghum. The importance of satellite DNA in influencing plant genome organization has been documented previously, and subtelomeric tandem arrays are characteristic of many plant genomes, raising the possibility that they play a role in telomere or genome stability (Mehrotra and Goyal, 2014; Padeken et al., 2015).The subtelomeric repeats STA1 and STA2 were located near the distal ends of most chromosomes. These sequences were identified as TRIM-like, although they lacked most of the sequence motifs found in TRIMs identified in other plants (Gao et al., 2016; Witte et al., 2001). The function of these subtelomeric repeats is unknown, although subtelomeric repeats

481 have been shown to be involved in bouquet formation and to facilitate the pairing of homologous
482 chromosomes during meiosis (Harper et al., 2004; Sadaie et al., 2003). A complete analysis of these
483 subtelomeric arrays will require additional long-read sequencing to fully characterize the size and
484 location of these subtelomeric repeats and to determine if they are present in all of the sorghum
485 chromosomes.

486 Comparison of whole genome sequences from 52 diverse sorghum genotypes to the v3 reference
487 genome sequence identified ~7.8M SNPs and ~1.9M indels. Large scale signals in the accumulation
488 of genetic variation were identified by signal processing techniques, and these may represent
489 signatures left by higher order organization. For example, elevated variant frequency was associated
490 with the wobble position in codons. Previous studies had documented elevated variant density in
491 euchromatic regions compared to pericentromeric regions of sorghum chromosomes and significant
492 variation in variant density within euchromatin when the genomes of different sorghum races were
493 compared (Evans et al., 2013). Genetic hitchhiking may be acting to reduce genetic variation in
494 regions of low recombination near centromeres (Barton, 2000). In this study, variant distributions
495 based on the analysis of 52 sorghum genomes were analyzed and found to contain large scale variant
496 distribution features that repeat every ~25 kbp. We had previously speculated that large scale
497 features like these could be generated by regional variation in recombination and repair, possibly due
498 to higher-order chromatin organization (Evans et al., 2013). In addition, the ability of DNA repair
499 machinery to access and correct mutations and selection pressures generated by functional properties
500 of the genome such as gene coding sequences across the gene rich distal arms of sorghum
501 chromosomes where rates of recombination are high could be influencing the accumulation of
502 variants (Evans et al., 2013; Mace et al., 2013; Makova and Hardison, 2015; Zheng et al., 2011).
503 Additional analyses should leverage wavelet transforms in addition to the discrete Fourier transform
504 to resolve problems associated non-stationary signals, as these genomic signals are likely non-
505 stationary in nature. Moreover, wavelet transform coefficients can be used to correlate multiple
506 features such as recombination and genetic variation (Spencer et al., 2006). The results from digital
507 signal processing approaches used to examine the sorghum genome indicate that additional
508 experimentation to annotate sorghum chromatin as well as higher order features like chromatin
509 interactions and nuclear lamina binding sites will be useful to better understand factors shaping the
510 landscape of the sorghum genome.

511 The RNA-seq transcriptome atlas reported here focused on the collection of tissue from growing and
512 fully developed roots, stems, leaves, panicles and seeds during development. Collection started with
513 seed germination, traversed the juvenile, vegetative and reproductive phases concluding with the
514 analysis of the transcriptome of dry seed. This transcriptome atlas complements prior RNA-seq data
515 collected from sorghum stems during 100 days of development that included the phase of sucrose
516 accumulation (McKinley et al., 2016), sorghum transcriptome responses to dehydration and ABA
517 (Dugas et al., 2011), dynamic changes in tiller bud transcriptomes modulated by PhyB (Kebrom and
518 Mullet, 2016), and an analysis of meristematic tissues, florets, and embryos (Olson et al., 2014). An
519 in depth description of the RNA-seq data is underway, however results described here show that the
520 atlas is of high quality and useful for the analysis of tissue and developmental states. The expression
521 of genes encoding kinases was found to differentiate transcriptome tissue states identified by PCA
522 analysis. Kinases are involved in plant development and tissue identity, and the transcriptome atlas
523 identified 47 genes encoding kinases whose transcript abundance broadly distinguishes between
524 tissue types. The kinase genes represent putative regulators of tissue identity in sorghum, and some
525 were previously characterized to influence plant development. Among the intersection of kinases
526 identified from the sorghum transcriptome atlas and those previously characterized in the literature
527 include kinases like WAK2, which is required for cell expansion during development by monitoring
528 pectin (Kohorn, 2015). TSL mediates RNAi silencing and may influence development (Uddin et al.,
529 2014). WNK4 and WNK6 were found to be regulated by the circadian clock and may be involved in
530 regulating flowering time (Nakamichi et al., 2002; Wang et al., 2008). ACR4 is associated with
531 maintenance of root stem cell identity in the RAM with CLV4, though ACR4 was not expressed in
532 roots in the transcriptome atlas (Stahl et al., 2013). ERL2 controls organ growth and flower
533 development via cell proliferation (Bemis et al., 2013; Shpak et al., 2004). YODA influences root
534 development through auxin up-regulation and cell division plane orientation (Smékalová et al.,
535 2014). These represent a small sampling of putative regulators of sorghum development, and thus the
536 sorghum transcriptome atlas represents a valuable resource with which to both annotate the sorghum
537 genome and to promote characterization of the gene regulatory networks underlying sorghum
538 development.

539

540 # 6 METHODS

541 ## 6.1 Genome assembly and improvement

320 regions of the version 1 sorghum reference genome assembly (Paterson et al., 2009) that contained a gene density greater than 2 genes per 100 kb were chosen for finishing. Finishing was performed by resequencing plasmid subclones and by walking on plasmid subclones or fosmids using custom primers. Small repeats in the sequence were resolved by transposon-hopping 8 kb plasmid clones, while 454 and Illumina based small insert libraries were used to improve resolution of simple sequence repeats. To fill large gaps, resolve large repeats, or to resolve chromosome duplications and extend into chromosome telomere regions, complete fosmid and BAC clones were shotgun sequenced and finished. The finished sequence was assembled, and each assembly was validated by an independent quality assessment. Finished regions were integrated by aligning the regions to the existing V1.0 assembly. 349 regions representing 344.4 Mbp of sequence were integrated in this manner.

A high-density genetic map generated from 437 recombinant inbred lines from a cross of BTx623 and IS3620C was used to improve the quality of the assembly and increase its coverage by integrating additional sequence scaffolds (Burow et al., 2011; Truong et al., 2014) into the 10 linkage groups. Scaffolds were broken if they contained a putative false join coincident with an area of low BAC/fosmid coverage. A total of 8 breaks were identified in the V1.0 release chromosomes, and an additional 7 previously unmapped scaffolds were integrated into the assembly in the appropriate location (Supplemental File S1). A 1.08 Mb region of the V1.0 chromosome 6 was moved to chromosome 7. 15 joins were made to form the final assembly containing 10 chromosomes capturing 655.2 Mb (97.1%) of the assembled sequence. Each join was padded with 10,000 Ns.

Homozygous variants identified from 110x of 2x250 (800 bp insert) Illumina fragments sequenced from the same DNA isolation as the original sequence were obtained and used to correct sequencing errors in the reference assembly. Reads were aligned to the integrated assembly and variants were called; variants that were called as homozygous were considered as candidates for correction in the reference assembly. A total of 1,942 (41% of called) homozygous SNPs and 1,432 (82% of called) homozygous indels were corrected in the process. SNPs and/or INDELs that were within 150bp of one another were not corrected. Additional information regarding methods of assembly and finishing are contained in Supplemental File S1.

6.2 Sample preparation and sequencing for transcriptome atlas and whole genome resequencing.

572 The reference line BTx623 was grown under 14 hour day greenhouse conditions in topsoil,
573 equivalent to native field soil from Brazos County, TX, to generate tissue for two separate
574 experiments: (1) a tissue by developmental stage timecourse, and (2) a nitrogen source study. For the
575 tissue by developmental stage timecourse, plants were harvested at the juvenile stage (8 DAE), the
576 vegetative stage (24 DAE), at floral initiation (44 DAE), at anthesis (65 DAE), and at grain maturity
577 (96 DAE) and leaf, root, stem and reproductive structures were flash frozen in liquid nitrogen. For
578 each tissue by stage combination, three biological replicates (i.e. three plants representing a single
579 condition) were harvested with the exception of the juvenile stage, for which a replicate was
580 represented by five plants instead of one to compensate for lower tissue abundance. For the nitrogen
581 source study, plants grown under differing nitrogen source regimes were harvested at 30 DAE, and
582 shoots and roots were flash frozen. For each tissue by condition, three biological replicates were
583 obtained. Additional details regarding harvested samples can be found in Supplemental Table S2 and
584 Supplemental Files S1 and S3.

585 Tissue was ground under liquid nitrogen and RNA was extracted using a Trizol-reagent based
586 extraction. Tissues with high levels of starch used a modified Trizol-reagent protocol (Li and Trick,
587 2005). Plate-based RNA sample prep was performed on the PerkinElmer Sciclone NGS robotic
588 liquid handling system using Illumina's TruSeq Stranded mRNA HT sample prep kit utilizing poly-A
589 selection of mRNA following the protocol outlined by Illumina in their user guide:
590 http://support.illumina.com/sequencing/sequencing_kits/truseq_stranded_mrna_ht_sample_prep_kit.
591 html, and with the following conditions: total RNA starting material was 1 ug per sample and 8
592 cycles of PCR was used for library amplification. The prepared libraries were then quantified by
593 qPCR using the Kapa SYBR Fast Illumina Library Quantification Kit (Kapa Biosystems) and run on
594 a Roche LightCycler 480 real-time PCR instrument. The quantified libraries were then prepared for
595 sequencing on the Illumina HiSeq sequencing platform utilizing a TruSeq paired-end cluster kit, v4,
596 and Illumina's cBot instrument to generate a clustered flowcell for sequencing. Sequencing of the
597 flowcell was performed on the Illumina HiSeq2500 sequencer using HiSeq TruSeq SBS sequencing
598 kits, v4, following a 2x150 indexed run recipe. Sequencing generated roughly 3.3 billion pairs of
599 sorghum paired-end read data.

600 Nine additional sorghum lines (100M, 80M, BTx623, BTx642, Hegari, IS3620C, SC170-6-17,
601 Standard Broomcorn, and Tx7000) were resequenced to supplement the 47 lines already available
602 (Mace et al., 2013; Zheng et al., 2011). Seeds were soaked in 20% bleach for 20 minutes and washed

603 extensively in distilled water for one hour. Seeds were germinated on water saturated germination
604 paper in a growth chamber (14 hr light; 30° C/10 hr dark; 24° C). Genomic DNA was isolated from
605 8-day old root tissue using a FastPrep DNA Extraction kit and FastPrep24 Instrument (MP
606 Biomedicals LLC, Solon, OH, USA), according to the manufacturer's specifications. DNA template
607 (350 bp average insert size) was prepared using a TruSeq® DNA PCR-Free LT Kit, according to the
608 manufacturer's directions. Paired-end sequencing (125 x 125 bases) was performed on an Illumina
609 HiSeq2500.

610 **6.3 Transcriptome Annotation**

611 The RNAseq reads were aligned to the updated reference assembly using GSNAP and assembled into
612 127,415 RNAseq transcripts with the PERTRAN pipeline (Shu et. al., unpublished). These
613 transcripts were combined with 209,835 ESTs to generate 111,994 transcript assemblies using
614 PASA. Loci were determined by transcript assembly alignments and/or EXONERATE alignments of
615 proteins from *Arabidopsis thaliana*, rice, maize or grape genomes. Gene models were predicted by
616 homology-based predictors, mainly FGENESH+, FGENESH_EST, and GenomeScan. The best
617 scored predictions for each locus were selected using multiple positive factors including EST and
618 protein support, and one negative factor: overlap with repeats. The selected gene predictions were
619 improved by PASA by adding UTRs, splicing correction, and adding alternative transcripts. Finally,
620 a homology analysis was performed on the PASA-improved models relative to the proteomes of
621 *Arabidopsis thaliana*, rice, maize and grape to identify high quality gene models and remove models
622 with extensive transposable element domains.

623 **6.4 Additional feature annotation, feature coverage, and periodicity analyses.**

624 Additional features were annotated in the sorghum genome, including repetitive sequence, genetic
625 variants, and nucleosome occupancy likelihoods. Repetitive sequence, including transposons and
626 SSRs, were annotated using both a de novo annotation and an annotation with existing libraries with
627 REPET v2.5; existing repetitive element libraries included the TIGR Plant Repeat Database and
628 RepBase (Bao et al., 2015; Flutre et al., 2011; Ouyang and Buell, 2004; Quesneville et al., 2005).
629 Genetic variants were called from sequence data for 56 sorghum resequenced sorghum samples
630 (Supplemental File S5). Processing of sequence reads to variant calls, including alignment to the Sbi3
631 reference genome, base recalibration, indel realignment, joint genotyping, and variant quality score
632 recalibration were performed using BWA v0.7.12 and GATK v3.3 and following the informed

pipeline of the RIG workflow (Auwera et al., 2013; DePristo et al., 2011; Li and Durbin, 2009; McCormick et al., 2015; McKenna et al., 2010). For examining variant accumulation at transcription start sites or coding sequence start sites, the v3.4 gene annotation was used. For all genes, the number of variants at each coordinate relative to the TSS or CDS were summed. For examining periodicity in genome-wide variant accumulation, the average number of variants in a 5,000 bp sliding window centered on the coordinate was determined, then scaled by a factor of 100 (i.e. number of SNPs per 50 base pairs averaged over 5,000 base pairs). To calculate nucleosome occupancy likelihoods, the support vector machine trained by Gupta et al. (2008) was used to calculate likelihoods of 50 bp sliding windows of primary sequence as in Fincher et al. (2013).

Periodicity of SNP accumulation or NOLs was performed using FFTPack within SciPy with the Fast Fourier Transformation (FFT). Genome-wide scans for periodicity were performed using a sliding window of the genome-wide variant accumulation (5,000 bp averages) and NOLs. The signal within a given window was transformed with the FFT, and windows meeting a set of criteria, including strength of a single frequency and a minimum number of cycles, were retained.

6.5 Characterization of STA1 and STA2

Sequence corresponding to STA1 and STA2 were identified initially by examining sequence underlying periodic NOLs. The STA1 and STA2 monomers were defined by finding the minimum complete repeat (~180 bp) using BLAST. The starts of the monomers were defined as the region of homology between STA1 and STA2, and for each, the consensus sequence of each monomer was determined by multiple sequence alignment of 9 different monomers representing a trio of tandem repeats from three different arrays on three different chromosome arms (Supplemental Figure S3 and Supplemental File S2) using multalin (Corpet, 1988). Extraction of sequence based on coordinates was facilitated using Biopieces (www.biopieces.org). Internal tandem direct repeats were identified using mreps and YASS (Kolpakov et al., 2003; Noé and Kucherov, 2005).

6.6 Gene expression analyses

Gene level read counts were obtained from RNA-seq reads and aligned individually to the version 3 assembly for each biological replicate. The FPKMs of three replicates of a condition were averaged to represent the sample. Per gene FPKMs were analyzed using the scikit-learn python package to perform dimensionality reduction and clustering (Pedregosa et al., 2011). Gene ontology analysis was performed using goatools Python package (Tang et al., 2015).

7 DATA ACCESS

The sorghum reference genome sequence and annotation are available from phytozome.jgi.doe.gov. The sequence has also been deposited in GenBank under accession number ABXC00000000. Sequence reads for the 56 resequenced lines are available in the in the National Center for Biotechnology Information Sequence Read Archive (NCBI SRA) under the IDs provided in Supplemental File S5; the 9 lines sequenced as part of this work are associated with BioProject PRJNA374837.

8 DISCLOSUE DECLARATION

The authors have no conflicts of interest to declare.

9 CONTRIBUTIONS

R.F.M. and S.K.T. performed downstream analyses (e.g. expression clustering, coverage analyses, periodicity analyses), transposon annotation, and linkage analyses. A.S. performed RNA-seq QC, read mapping and expression analyses. S.S. performed gene annotation (gene set version 3.1). J.J, D.S., and J.G. performed genome assembly and finishing (genome version 3.0). M.K. and M.A. performed sorghum transcriptome atlas sequencing. R.F.M., S.K.T., B.W., B.M., and A.M. prepared transcriptome atlas samples. D.M. performed resequencing of selected sorghum lines. J.G., J.S., and J.M. conceived and provided project management. R.F.M., S.K.T., and J.M. wrote the manuscript. All authors reviewed and approved of the manuscript.

10 ACKNOWLEDGEMENTS

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This work was funded in part by the DOE Great Lakes Bioenergy Research Center (DOE Office of

689 Science BER DE-FC02-07ER64494), and the U.S. Department of Energy grants no. DE-AR0000596
690 and DE-SC0012629.

691

11 REFERENCES

Abdel-Ghany, S. E., Hamilton, M., Jacobi, J. L., Ngam, P., Devitt, N., Schilkey, F., Ben-Hur, A., and Reddy, A. S. (2016). A survey of the sorghum transcriptome using single-molecule long reads. *Nature Communications* **7**.

Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., and Thibault, J. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, 11.10. 1-11.10. 33.

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 1.

Barton, N. H. (2000). Genetic hitchhiking. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **355**, 1553-1562.

Bemis, S. M., Lee, J. S., Shpak, E. D., and Torii, K. U. (2013). Regulation of floral patterning and organ identity by Arabidopsis ERECTA-family receptor kinase genes. *Journal of experimental botany* **64**, 5323-5333.

Bennetzen, J. L., Schmutz, J., Wang, H., Percifield, R., Hawkins, J., Pontaroli, A. C., Estep, M., Feng, L., Vaughn, J. N., and Grimwood, J. (2012). Reference genome sequence of the model plant Setaria. *Nature biotechnology* **30**, 555-561.

Bonev, B., and Cavalli, G. (2016). Organization and function of the 3D genome. *Nature Reviews Genetics* **17**, 661-678.

Bowers, J. E., Abbey, C., Anderson, S., Chang, C., Draye, X., Hoppe, A. H., Jessup, R., Lemke, C., Lennington, J., and Li, Z. (2003). A high-density genetic recombination map of sequence-tagged sites for sorghum, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. *Genetics* **165**, 367-386.

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G. L., D'Amore, R., Allen, A. M., McKenzie, N., Kramer, M., Kerhornou, A., and Bolser, D. (2012). Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**, 705-710.

Brenton, Z. W., Cooper, E. A., Myers, M. T., Boyles, R. E., Shakoor, N., Zielinski, K. J., Rauh, B. L., Bridges, W. C., Morris, G. P., and Kresovich, S. (2016). A genomic resource for the development, improvement, and exploitation of sorghum for bioenergy. *Genetics* **204**, 21-33.

Burow, G. B., Klein, R. R., Franks, C. D., Klein, P. E., Schertz, K. F., Pederson, G. A., Xin, Z., and Burke, J. J. (2011). Registration of the BTx623/IS3620C Recombinant Inbred Mapping Population of Sorghum. *Journal of Plant Registrations* **5**, 141-145.

Consortium, E. P. (2012a). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74.

Consortium, I. B. G. S. (2012b). A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491**, 711-716.

Consortium, I. H. G. S. (2004). Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945.

Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic acids research* **16**, 10881-10890.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., and Hanna, M. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498.

Dugas, D. V., Monaco, M. K., Olson, A., Klein, R. R., Kumari, S., Ware, D., and Klein, P. E. (2011). Functional annotation of the transcriptome of Sorghum bicolor in response to osmotic stress and abscisic acid. *BMC genomics* **12**, 514.

Evans, J., McCormick, R. F., Morishige, D., Olson, S. N., Weers, B., Hilley, J., Klein, P., Rooney, W., and Mullet, J. (2013). Extensive variation in the density and distribution of DNA polymorphism in sorghum genomes. *PloS one* **8**, e79192.

Fincher, J. A., Vera, D. L., Hughes, D. D., McGinnis, K. M., Dennis, J. H., and Bass, H. W. (2013). Genome-wide prediction of nucleosome occupancy in maize reveals plant chromatin structural features at genes and other elements at multiple scales. *Plant physiology* **162**, 1127-1141.

Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. (2011). Considering transposable element diversification in de novo annotation approaches. *PloS one* **6**, e16526.

Gao, D., Li, Y., Do Kim, K., Abernathy, B., and Jackson, S. A. (2016). Landscape and evolutionary dynamics of terminal repeat retrotransposons in miniature in plant genomes. *Genome biology* **17**, 1.

Gupta, S., Dennis, J., Thurman, R. E., Kingston, R., Stamatoyannopoulos, J. A., and Noble, W. S. (2008). Predicting human nucleosome occupancy from primary sequence. *PLoS Comput Biol* **4**, e1000134.

Harper, L., Golubovskaya, I., and Cande, W. Z. (2004). A bouquet of chromosomes. *Journal of Cell Science* **117**, 4025-4032.

Higasa, K., and Hayashi, K. (2006). Periodicity of SNP distribution around transcription start sites. *BMC genomics* **7**, 66.

Hilley, J., Truong, S., Olson, S., Morishige, D., and Mullet, J. (2016). Identification of Dw1, a regulator of sorghum stem internode length. *PloS one* **11**, e0151271.

Hirsch, C. N., Foerster, J. M., Johnson, J. M., Sekhon, R. S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M. A., and Barry, K. (2014). Insights into the maize pan-genome and pan-transcriptome. *The Plant Cell* **26**, 121-135.

Hoang-Tang, Dube, S. K., Liang, G. H., and Kung, S.-D. (1991). Possible repetitive DNA markers for Eusorghum and Parasorghum and their potential use in examining phylogenetic hypotheses on the origin of Sorghum species. *Genome* **34**, 241-250.

Hodgkinson, A., and Eyre-Walker, A. (2011). Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics* **12**, 756-766.

Kebrom, T. H., and Mullet, J. E. (2016). Transcriptome profiling of tiller buds provides new insights into PhyB regulation of tillering and indeterminate growth in sorghum. *Plant physiology* **170**, 2232-2250.

Kim, J.-S., Klein, P. E., Klein, R. R., Price, H. J., Mullet, J. E., and Stelly, D. M. (2005). Chromosome identification and nomenclature of Sorghum bicolor. *Genetics* **169**, 1169-1173.

Klein, P. E., Klein, R. R., Cartinhour, S. W., Ulanch, P. E., Dong, J., Obert, J. A., Morishige, D. T., Schlueter, S. D., Childs, K. L., and Ale, M. (2000). A high-throughput AFLP-based method for constructing integrated genetic and physical maps: progress toward a sorghum genome map. *Genome Research* **10**, 789-807.

Kohorn, B. D. (2015). The state of cell wall pectin monitored by wall associated kinases: A model. *Plant signaling & behavior* **10**, e1035854.

Kolpakov, R., Bana, G., and Kucherov, G. (2003). mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic acids research* **31**, 3672-3678.

Kromdijk, J., Głowacka, K., Leonelli, L., Gabilly, S. T., Iwai, M., Niyogi, K. K., and Long, S. P. (2016). Improving photosynthesis and crop productivity by accelerating recovery from photoprotection. *Science* **354**, 857-861.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., and FitzHugh, W. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754-1760.

Li, Z., and Trick, H. N. (2005). Rapid method for high-quality RNA isolation from seed endosperm containing high levels of starch. *Biotechniques* **38**, 872.

Liu, M.-J., Seddon, A. E., Tsai, Z. T.-Y., Major, I. T., Floer, M., Howe, G. A., and Shiu, S.-H. (2015). Determinants of nucleosome positioning and their influence on plant gene expression. *Genome research* **25**, 1182-1195.

Mace, E. S., Tai, S., Gilding, E. K., Li, Y., Prentis, P. J., Bian, L., Campbell, B. C., Hu, W., Innes, D. J., and Han, X. (2013). Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nature communications* **4**.

789 Makova, K. D., and Hardison, R. C. (2015). The effects of chromatin organization on variation in mutation
790 rates in the genome. *Nature Reviews Genetics* **16**, 213-223.
791 McCormick, R. F., Truong, S. K., and Mullet, J. E. (2015). RIG: Recalibration and Interrelation of Genomic
792 Sequence Data with the GATK. *G3-Genes Genomes Genetics* **5**, 655-665.
793 McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D.,
794 Gabriel, S., and Daly, M. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing
795 next-generation DNA sequencing data. *Genome research* **20**, 1297-1303.
796 McKinley, B., Rooney, W., Wilkerson, C., and Mullet, J. (2016). Dynamics of biomass partitioning, stem gene
797 expression, cell wall biosynthesis, and sucrose accumulation during development of Sorghum bicolor.
798 *The Plant Journal* **88**, 662-680.
799 Mehrotra, S., and Goyal, V. (2014). Repetitive sequences in plant nuclear DNA: types, distribution, evolution
800 and function. *Genomics, proteomics & bioinformatics* **12**, 164-171.
801 Mickelbart, M. V., Hasegawa, P. M., and Bailey-Serres, J. (2015). Genetic mechanisms of abiotic stress
802 tolerance that translate to crop yield stability. *Nature Reviews Genetics* **16**, 237-251.
803 Mondal, S., Rutkoski, J. E., Velu, G., Singh, P. K., Crespo-Herrera, L. A., Guzman, C. G., Bhavani, S., Lan, C., He,
804 X., and Singh, R. P. (2016). Harnessing diversity in wheat to enhance grain yield, climate resilience,
805 disease and insect pest resistance and nutrition through conventional and modern breeding
806 approaches. *Frontiers in Plant Science* **7**, 991.
807 Morishige, D. T., Klein, P. E., Hilley, J. L., Sahraeian, S. M. E., Sharma, A., and Mullet, J. E. (2013). Digital
808 genotyping of sorghum - a diverse plant species with a large repeat-rich genome. *Bmc Genomics* **14**.
809 Morris, G. P., Ramu, P., Deshpande, S. P., Hash, C. T., Shah, T., Upadhyaya, H. D., Riera-Lizarazu, O., Brown, P.
810 J., Acharya, C. B., and Mitchell, S. E. (2013). Population genomic and genome-wide association
811 studies of agroclimatic traits in sorghum. *Proceedings of the National Academy of Sciences* **110**, 453-
812 458.
813 Mullet, J., Morishige, D., McCormick, R., Truong, S., Hilley, J., McKinley, B., Anderson, R., Olson, S. N., and
814 Rooney, W. (2014). Energy Sorghum-a genetic model for the design of C-4 grass bioenergy crops.
815 *Journal of Experimental Botany* **65**, 3479-3489.
816 Murphy, R. L., Klein, R. R., Morishige, D. T., Brady, J. A., Rooney, W. L., Miller, F. R., Dugas, D. V., Klein, P. E.,
817 and Mullet, J. E. (2011). Coincident light and clock regulation of pseudoresponse regulator protein 37
818 (PRR37) controls photoperiodic flowering in sorghum. *Proceedings of the National Academy of*
819 *Sciences* **108**, 16469-16474.
820 Nakamichi, N., Murakami-Kojima, M., Sato, E., KISHI, Y., YAMASHINO, T., and MIZUNO, T. (2002). Compilation
821 and characterization of a novel WNK family of protein kinases in Arabiodpsis thaliana with reference
822 to circadian rhythms. *Bioscience, biotechnology, and biochemistry* **66**, 2429-2436.
823 Noé, L., and Kucherov, G. (2005). YASS: enhancing the sensitivity of DNA similarity search. *Nucleic acids*
824 *research* **33**, W540-W543.
825 Olson, A., Klein, R. R., Dugas, D. V., Lu, Z., Regulski, M., Klein, P. E., and Ware, D. (2014). Expanding and
826 vetting gene annotations through transcriptome and methylome sequencing. *The Plant Genome* **7**.
827 Ort, D. R., Merchant, S. S., Alric, J., Barkan, A., Blankenship, R. E., Bock, R., Croce, R., Hanson, M. R., Hibberd,
828 J. M., Long, S. P., Moore, T. A., Moroney, J., Niyogi, K. K., Parry, M. A. J., Peralta-Yahya, P. P., Prince,
829 R. C., Redding, K. E., Spalding, M. H., van Wijk, K. J., Vermaas, W. F. J., von Caemmerer, S., Weber, A.
830 P. M., Yeates, T. O., Yuan, J. S., and Zhu, X. G. (2015). Redesigning photosynthesis to sustainably meet
831 global food and bioenergy demand. *Proceedings of the National Academy of Sciences of the United*
832 *States of America* **112**, 8529-8536.
833 Ouyang, S., and Buell, C. R. (2004). The TIGR Plant Repeat Databases: a collective resource for the
834 identification of repetitive sequences in plants. *Nucleic acids research* **32**, D360-D363.
835 Padeken, J., Zeller, P., and Gasser, S. M. (2015). Repeat DNA in genome organization and stability. *Current*
836 *opinion in genetics & development* **31**, 12-19.
837 Park, S.-Y., Peterson, F. C., Mosquna, A., Yao, J., Volkman, B. F., and Cutler, S. R. (2015). Agrochemical control
838 of plant water use using engineered abscisic acid receptors. *Nature* **520**, 545-548.

839 Participants, S. G. P. W. (2005). Toward sequencing the sorghum genome. A US National Science Foundation-
840 sponsored workshop report. *Plant Physiology*, 1898-1902.
841 Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten,
842 U., Mitros, T., and Poliakov, A. (2009). The Sorghum bicolor genome and the diversification of
843 grasses. *Nature* **457**, 551-556.
844 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P.,
845 Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine*
846 *Learning Research* **12**, 2825-2830.
847 Price, H. J., Dillon, S. L., Hodnett, G., Rooney, W. L., Ross, L., and Johnston, J. S. (2005). Genome evolution in
848 the genus Sorghum (Poaceae). *Annals of Botany* **95**, 219-227.
849 Quesneville, H., Bergman, C. M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., and Anxolabehere, D.
850 (2005). Combined evidence annotation of transposable elements in genome sequences. *PLoS*
851 *Comput Biol* **1**, e22.
852 Rosenbloom, K. R., Sloan, C. A., Malladi, V. S., Dreszer, T. R., Learned, K., Kirkup, V. M., Wong, M. C.,
853 Maddren, M., Fang, R., and Heitner, S. G. (2013). ENCODE data in the UCSC Genome Browser: year 5
854 update. *Nucleic acids research* **41**, D56-D63.
855 Sadaie, M., Naito, T., and Ishikawa, F. (2003). Stable inheritance of telomere chromatin structure and
856 function in the absence of telomeric repeats. *Genes & development* **17**, 2271-2282.
857 Sanborn, A. L., Rao, S. S., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., Bochkov, I. D., Chinnappan,
858 D., Cutkosky, A., and Li, J. (2015). Chromatin extrusion explains key features of loop and domain
859 formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*
860 **112**, E6456-E6465.
861 Sasaki, S., Mello, C. C., Shimada, A., Nakatani, Y., Hashimoto, S.-i., Ogawa, M., Matsushima, K., Gu, S. G.,
862 Kasahara, M., and Ahsan, B. (2009). Chromatin-associated periodicity in genetic variation
863 downstream of transcriptional start sites. *Science* **323**, 401-404.
864 Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., and
865 Graves, T. A. (2009). The B73 maize genome: complexity, diversity, and dynamics. *science* **326**, 1112-
866 1115.
867 Sekhon, R. S., Briskine, R., Hirsch, C. N., Myers, C. L., Springer, N. M., Buell, C. R., de Leon, N., and Kaeppler, S.
868 M. (2013). Maize gene atlas developed by RNA sequencing and comparative evaluation of
869 transcriptomes based on RNA sequencing and microarrays. *PLoS One* **8**, e61005.
870 Sekhon, R. S., Lin, H., Childs, K. L., Hansey, C. N., Buell, C. R., de Leon, N., and Kaeppler, S. M. (2011). Genome-
871 wide atlas of transcription during maize development. *The Plant Journal* **66**, 553-563.
872 Shakoor, N., Nair, R., Crasta, O., Morris, G., Feltus, A., and Kresovich, S. (2014). A Sorghum bicolor expression
873 atlas reveals dynamic genotype-specific expression profiles for vegetative tissues of grain, sweet and
874 bioenergy sorghums. *BMC plant biology* **14**, 1.
875 Shpak, E. D., Berthiaume, C. T., Hill, E. J., and Torii, K. U. (2004). Synergistic interaction of three ERECTA-
876 family receptor-like kinases controls Arabidopsis organ growth and flower development by
877 promoting cell proliferation. *Development* **131**, 1491-1501.
878 Smékalová, V., Luptovčiak, I., Komis, G., Šamajová, O., Ovečka, M., Doskočilová, A., Takáč, T., Vadovič, P.,
879 Novák, O., and Pechan, T. (2014). Involvement of YODA and mitogen activated protein kinase 6 in
880 Arabidopsis post-embryogenic root development through auxin up-regulation and cell division plane
881 orientation. *New Phytologist* **203**, 1175-1193.
882 Spencer, C. C., Deloukas, P., Hunt, S., Mullikin, J., Myers, S., Silverman, B., Donnelly, P., Bentley, D., and
883 McVean, G. (2006). The influence of recombination on human genetic diversity. *PLoS Genet* **2**, e148.
884 Stahl, Y., Grabowski, S., Bleckmann, A., Kühnemuth, R., Weidtkamp-Peters, S., Pinto, K. G., Kirschner, G. K.,
885 Schmid, J. B., Wink, R. H., and Hülsewede, A. (2013). Moderation of Arabidopsis root stemness by
886 CLAVATA1 and ARABIDOPSIS CRINKLY4 receptor kinase complexes. *Current Biology* **23**, 362-371.
887 Tang, H., Klopfenstein, D., Pederson, B., Flick, P., Sato, K., Ramirez, F., Yunes, J., and Mungall, C. (2015).
888 GOATOOLS: Tools for Gene Ontology. *Zendo*.

889 Technow, F., Messina, C. D., Totir, L. R., and Cooper, M. (2015). Integrating Crop Growth Models with Whole
890 Genome Prediction through Approximate Bayesian Computation. *Plos One* **10**.
891 Tolstorukov, M. Y., Volfovsky, N., Stephens, R. M., and Park, P. J. (2011). Impact of chromatin structure on
892 sequence variability in the human genome. *Nature structural & molecular biology* **18**, 510-515.
893 Truong, S. K., McCormick, R. F., Morishige, D. T., and Mullet, J. E. (2014). Resolution of Genetic Map
894 Expansion Caused by Excess Heterozygosity in Plant Recombinant Inbred Populations. *G3-Genes*
895 *Genomes Genetics* **4**, 1963-1969.
896 Uddin, M. N., Dunoyer, P., Schott, G., Akhter, S., Shi, C., Lucas, W. J., Voinnet, O., and Kim, J.-Y. (2014). The
897 protein kinase TOUSLED facilitates RNAi in Arabidopsis. *Nucleic acids research* **42**, 7971-7980.
898 VanBuren, R., Bryant, D., Edger, P. P., Tang, H., Burgess, D., Challabathula, D., Spittle, K., Hall, R., Gu, J., and
899 Lyons, E. (2015). Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum.
900 *Nature*.
901 Vogel, J. P., Garvin, D. F., Mockler, T. C., Schmutz, J., Rokhsar, D., Bevan, M. W., Barry, K., Lucas, S., Harmon-
902 Smith, M., and Lail, K. (2010). Genome sequencing and analysis of the model grass Brachypodium
903 distachyon. *Nature* **463**, 763-768.
904 Voytas, D. F. (2013). Plant genome engineering with sequence-specific nucleases. *Plant Biology* **64**, 327.
905 Wang, L., Xie, W., Chen, Y., Tang, W., Yang, J., Ye, R., Liu, L., Lin, Y., Xu, C., and Xiao, J. (2010). A dynamic gene
906 expression atlas covering the entire life cycle of rice. *The Plant Journal* **61**, 752-766.
907 Wang, Y., Liu, K., Liao, H., Zhuang, C., Ma, H., and Yan, X. (2008). The plant WNK gene family and regulation of
908 flowering time in Arabidopsis. *Plant Biology* **10**, 548-562.
909 Witte, C.-P., Le, Q. H., Bureau, T., and Kumar, A. (2001). Terminal-repeat retrotransposons in miniature
910 (TRIM) are involved in restructuring plant genomes. *Proceedings of the National Academy of Sciences*
911 **98**, 13778-13783.
912 Zhang, G., Liu, X., Quan, Z., Cheng, S., Xu, X., Pan, S., Xie, M., Zeng, P., Yue, Z., and Wang, W. (2012). Genome
913 sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential.
914 *Nature biotechnology* **30**, 549-554.
915 Zheng, L.-Y., Guo, X.-S., He, B., Sun, L.-J., Peng, Y., Dong, S.-S., Liu, T.-F., Jiang, S., Ramachandran, S., and Liu,
916 C.-M. (2011). Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum
917 bicolor). *Genome biology* **12**, 1.
918 Zwick, M., Islam-Faridi, M., Zhang, H., Hodnett, G., Gomez, M., Kim, J., Price, H., and Stelly, D. (2000).
919 Distribution and sequence analysis of the centromere-associated repetitive element CEN38 of
920 Sorghum bicolor (Poaceae). *American Journal of Botany* **87**, 1757-1764.

921