Title

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



Journal

The Plant Journal, 93(2)

ISSN

0960-7412

Authors

McCormick, Ryan F

Truong, Sandra K

Sreedasyam, Avinash

et al.

Publication Date

2018

DOI

10.1111/tpj.13781

Peer reviewed

the plant journal

SEB Society for Experimental Biology

*The Plant Journal* (2018) **93,** 338–354 doi: 10.1111/tpj.13781

RESOURCE

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

**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 D. Weers[2], Brian McKinley[2], Ashley Mattison[1,2], Daryl T. Morishige[2], Jane Grimwood[3,4], Jeremy Schmutz[3,4] and John E. 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, and*
[4]*Department of Energy, Joint Genome Institute, Walnut Creek, CA 94598, USA*

Received 5 April 2017; revised 5 November 2017; accepted 14 November 2017; published online 21 November 2017.
*For correspondence (e-mail jmullet@tamu.edu).

## SUMMARY

***Sorghum bicolor*** **is a drought tolerant C4 grass used for the production of grain, forage, sugar, and lignocellulosic biomass and a genetic model for C4 grasses due to its relatively small genome (approximately 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. Subtelomeric 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 more than 50 resequenced genomes from diverse sorghum genotypes to the reference genome identified approximately 7.4 M single nucleotide polymorphisms (SNPs) and 1.9 M indels. Large-scale variant features in euchromatin were identified with periodicities of approximately 25 kbp. A transcriptome atlas of gene expression was constructed from 47 RNA-seq profiles of growing and developed tissues of the major plant organs (roots, leaves, stems, panicles, and 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.**

**Keywords: discrete Fourier transform, gene annotation, genetic variation, genome assembly, kinase, nucleosome occupancy, reference genome, satellite DNA, *Sorghum bicolor*.**

## 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 (Voytas, 2013; Mullet *et al.*, 2014; Mickelbart *et al.*, 2015; Ort *et al.*, 2015; Park *et al.*, 2015; Technow *et al.*, 2015; Kromdijk *et al.*, 2016; Mondal *et al.*, 2016). Progress toward the genetic improvement of plants is

promoted by the availability of foundational genetic and genomic resources. Because of this, we improved the *S. bicolor* reference genome sequence assembly 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 approximately 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). An *S. 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 (International Barley Genome Sequencing Consortium, 2012), and the desiccation tolerant *Oropetium thomaeum* (Van Buren *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 (Zheng *et al.*, 2011; Evans *et al.*, 2013; Mace *et al.*, 2013). 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 (Lander *et al.*, 2001; International Human Genome Sequencing Consortium, 2004, ENCODE Project Consortium, 2012, 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 (Wang *et al.*, 2010; Sekhon *et al.*, 2011, 2013). In sorghum, microarray-based expression profiling and RNA-seq have also been used to examine transcriptome dynamics in different sorghum genotypes, tissues, and responses to hormones and the environment (Shakoor *et al.*, 2014; Abdel-Ghany *et al.*, 2016; Kebrom and Mullet, 2016; McKinley *et al.*, 2016). 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 (Hodgkinson and Eyre-Walker, 2011; Tolstorukov *et al.*, 2011; Evans *et al.*, 2013; Makova and Hardison, 2015). 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). As 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 (Gupta *et al.*, 2008; Fincher *et al.*, 2013). Given that eukaryotic genomes are organized into higher order topologically associating domains and the influence of nucleosome occupancy on the accumulation of genetic variation, the possibility that larger chromatin domains influence the genome in a similar manner in plants also exists (Bonev and Cavalli, 2016). As such, we explored the basis of genetic variation accumulation in the sorghum genome using digital signal processing techniques.

## RESULTS

### Genome assembly and improvement

Version 1 of the sorghum BTx623 reference genome assembly incorporated 625.6 Mbp of genomic sequence into 10 pseudomolecules corresponding to the 10 sorghum

 

chromosomes by combining data from whole-genome shotgun sequencing and targeted sequencing of BACs and fosmids using paired-end Sanger sequencing. An error rate of <1 per 10 kbp was estimated based on Sanger sequencing of BACs (Paterson *et al.*, 2009). Version 2 of the sorghum reference genome assembly was publicly released without a corresponding publication; as such, all comparisons here are made relative to version 1.

In this study, version 1 of the sorghum reference genome was refined by deep whole-genome short read sequencing (110×) and targeted finishing of gene-dense regions of the genome (greater than 2 genes per 100 kbp) using primer walking via Sanger sequencing and shotgun sequencing of plasmid subclones, fosmid, and BAC clones (Data S1). These finished regions were assembled and hand-curated (representing 344.4 Mbp), mapped back to the v1 assembly, and then incorporated into the v1 assembly, adding a total of 4.96 Mbp to the gene rich portion of the genome assembly during the finishing process. To improve the order and coverage of the reference genome, a high-density genetic map based on approximately 10 000 markers genotyped in a 437-line recombinant inbred mapping population derived from the sorghum lines BTx623 and IS3620C was used to map the location and integrate seven sequence contigs spanning 24.64 Mbp into the v1 reference genome chromosomes (Truong *et al.*, 2014). Furthermore, the genetic map identified a 1.08 Mbp region that was previously assembled into chromosome 6, but markers within the region were not linked to flanking regions on chromosome 6 and were instead tightly linked with markers on chromosome 7 (Figure 1). This assembly error in version 1 is corrected in version 3.

Due to integration of additional sequence during finishing and of previously unplaced sequence 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 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).

### Annotation of genes and other features in the sorghum genome

The version 3 assembly was annotated for a number of feature types, including genes, repetitive elements, genetic variation, and primary sequence-based nucleosome occupancy predictions (Figure 2; 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

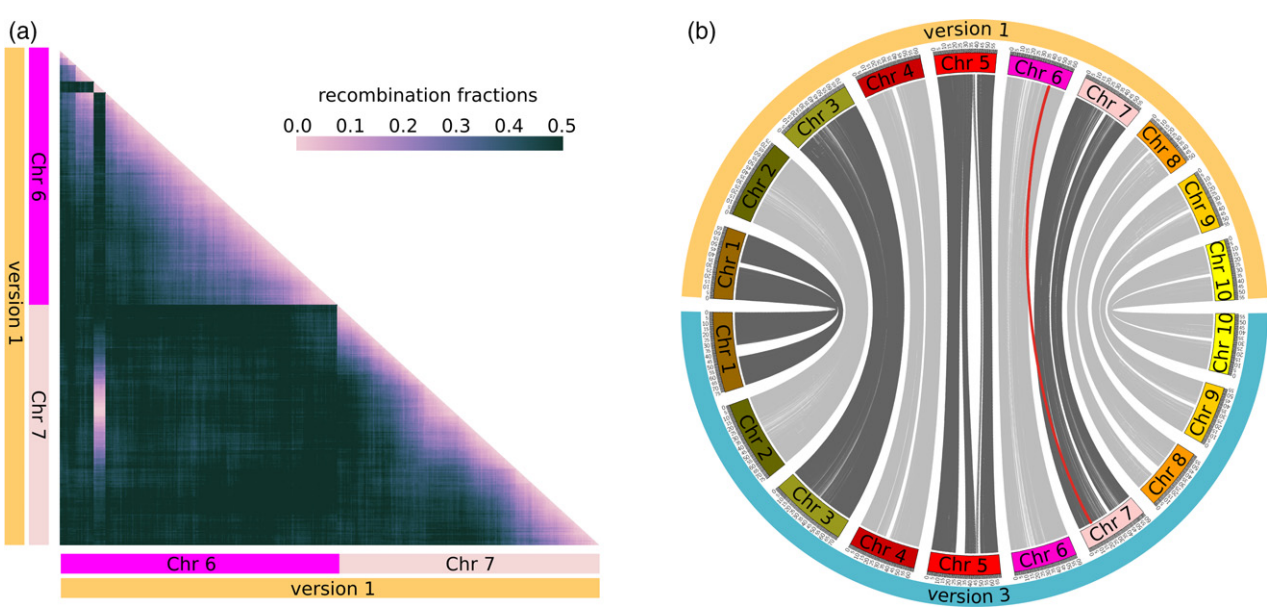


**Figure 1.** Correction of misassembled region in the version 1 sorghum reference genome assembly and integration of new sequence.
(a) Recombination fractions of markers in the BTx623 × IS3620C sorghum recombinant inbred line (RIL) population ordered by physical position relative to the version 1 reference 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 (Data S1).

**Table 1** Summary statistics for the *Sorghum bicolor* reference genome. 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 statistics | |
|---|---|---|
| | Version 1 | Version 3 |
| Number of pseudomolecules | 10 | 10 |
| Number of contigs | 6929 | 2688 |
| 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 |

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 *S. 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). A small number (175) of genes in v1.4 were not supported, and these were not included in the v3.1 gene set. More importantly, 6 989 new gene models supported by expression data and/or homology to known genes were added to the sorghum gene annotation. The new gene models were enriched in gene ontologies related to defense responses, photosynthesis, DNA integration, sulfate assimilation and KEGG pathways related to photosynthesis and oxidative phosphorylation. The median and mean gene size in v3.1 increased to 1 600 and 1 835, from 1 336 and 1 473 in v1.4, respectively, due primarily to improved annotation of exons (Figure S3). 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. Repetitive elements in the sorghum genome were annotated using a *de novo* repetitive element annotation pipeline in conjunction with existing repetitive element libraries (Ouyang and Buell, 2004; Quesneville *et al.*, 2005; Flutre *et al.*, 2011; Bao *et al.*, 2015). 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 was 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 1 Mbp sliding windows (Figure 2; 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 (Paterson *et al.*, 2009; Evans *et al.*, 2013). The accumulation of genetic variation in sorghum accessions was examined by aligning and comparing reads from 56 resequenced sorghum genotypes to the v3 genome sequence. Two *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 *S. 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 SVM 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 (Gupta *et al.*, 2008; Fincher *et al.*, 2013). 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. Moreover, because the SVM was trained on human data and validated in maize, sequences with high scores should be interpreted as evolutionarily conserved predispositions to nucleosome occupancy rather than definitive, species-specific occupancy. The predicted NOLs for sorghum are similar to maize in that the distributions vary across each chromosome in a manner not directly related to gene or repeat density across each chromosome (Figure 2; Figures S1 and S2). It is also notable that the presence of uncalled bases (i.e. Ns) has a pronounced effect on all of the annotations referred to above by either precluding annotation of features and causing low feature density, or by causing a fixed numeric assignment as in the case of NOLs. As such, the location of uncalled bases influences the interpretation of these feature distributions (Figure 2; Figures S1 and S2).

 

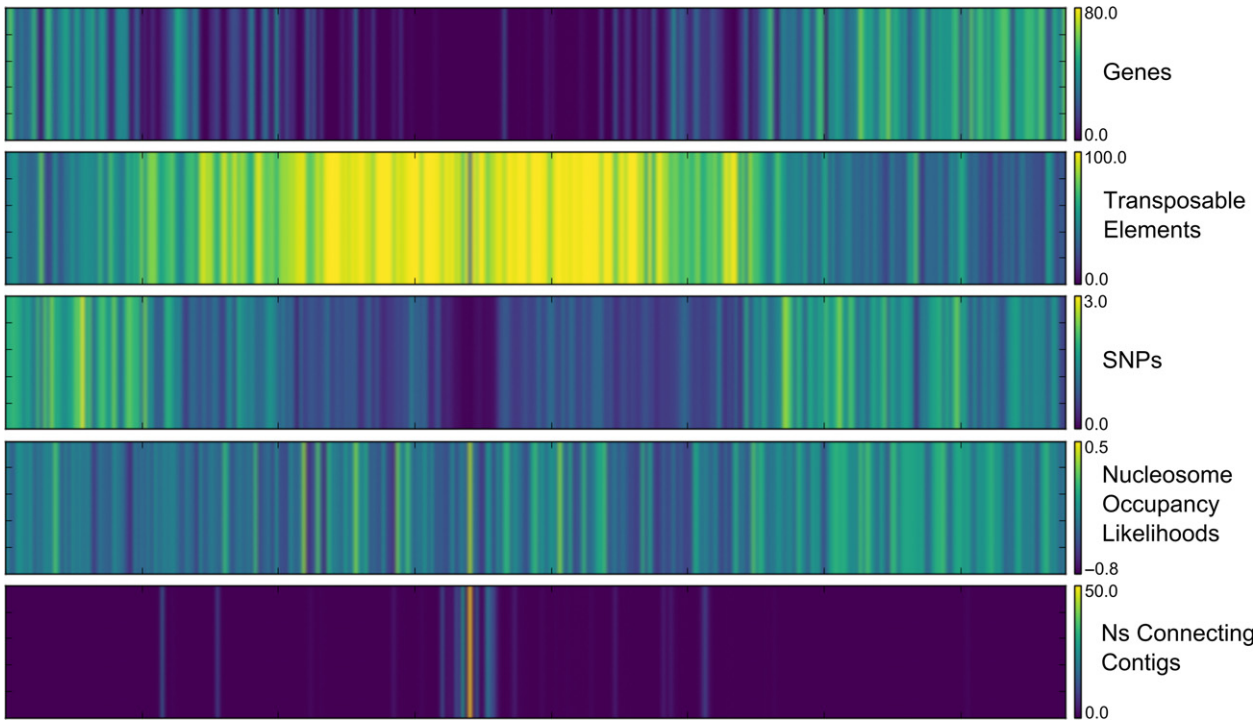


**Figure 2.** Feature densities and score averages across chromosome 2 of the sorghum genome.
Color map displaying the average densities of multiple features across chromosome 2 of the sorghum genome, including annotated genes, transposable elements, single nucleotide polymorphisms, nucleosome occupancy likelihoods (NOLs), and uncalled bases (Ns) connecting contigs in the assembly. For features other than NOLs, scales represent the percentage of nucleotides annotated as the feature. For NOLs, the scale represents the estimated likelihood of occupancy where positive values predict the presence of a nucleosome and negative values predict the absence of a nucleosome. Maps for all 10 chromosomes are depicted in Figures S1 and S2.

### Periodicity of features 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 distribution 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 (DFT) was used to examine periodicities in the accumulation of genetic variation and NOLs 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 sequence start sites, but not upstream (Figure 3).

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 accumulation of genetic variation and NOLs around TSSs 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 approximately 150 bp downstream of these sites. Nucleosome occupancy predictions also did

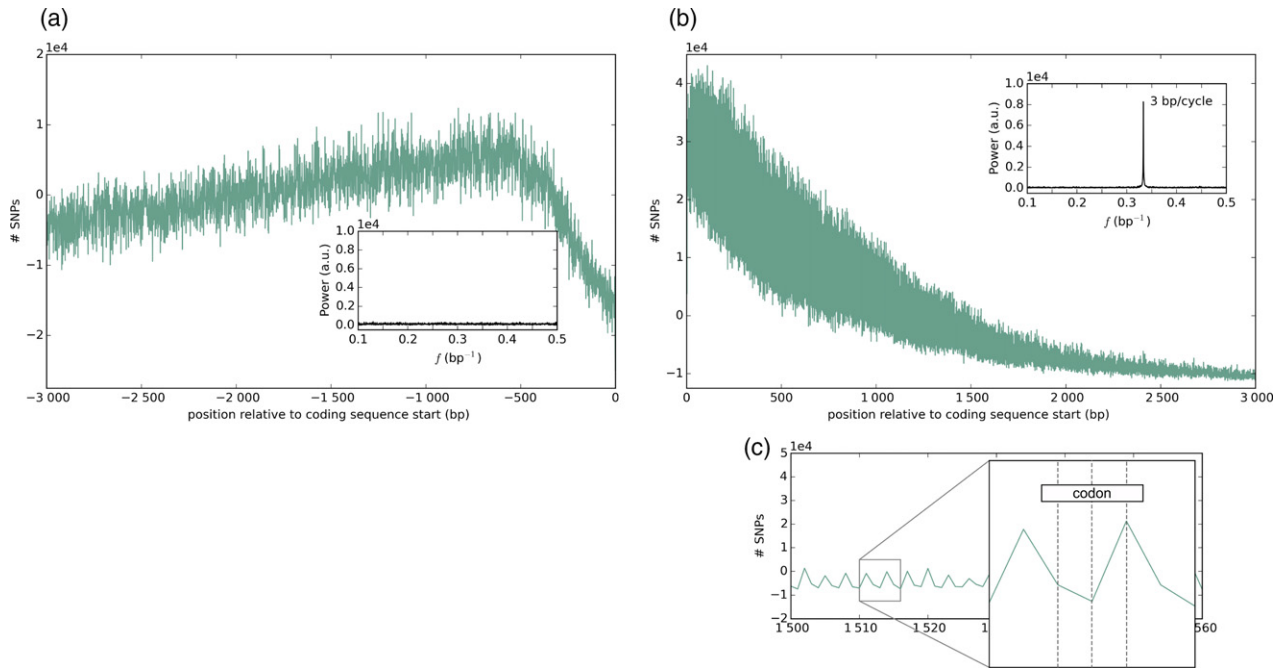


**Figure 3.** Functional properties of the sorghum genome leave periodic signatures that can be identified using signal processing techniques.
(a, b) Due to the degeneracy of the genetic code, relaxed selection at the wobble base in codons causes SNPs to accumulate with a periodicity of 3 bp downstream 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 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 upstream (a) or downstream (b) window. The apparent rapid decline of SNP abundance in panel A between −500 bp and 0 bp is putatively a consequence of increased selection at promoters and 5′ UTRs. The apparent decline of SNP abundance in panel B occurs because CDS that have lengths of less than 3 000 bp were considered to have 0 SNPs between their end position and 3 000 bp.

not predict a periodic arrangement of nucleosomes downstream of TSSs.

Nucleosome-scale periods of 180 bp are present in NOL 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 5(b, c)) (Klein *et al.*, 2000).

As the SVM used for NOL 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 5(a)). The two monomers are referred to as subtelomeric tandemly arrayed (STA)1 and STA2 here for brevity.

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 3 236; 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 (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 that included sorghum

 

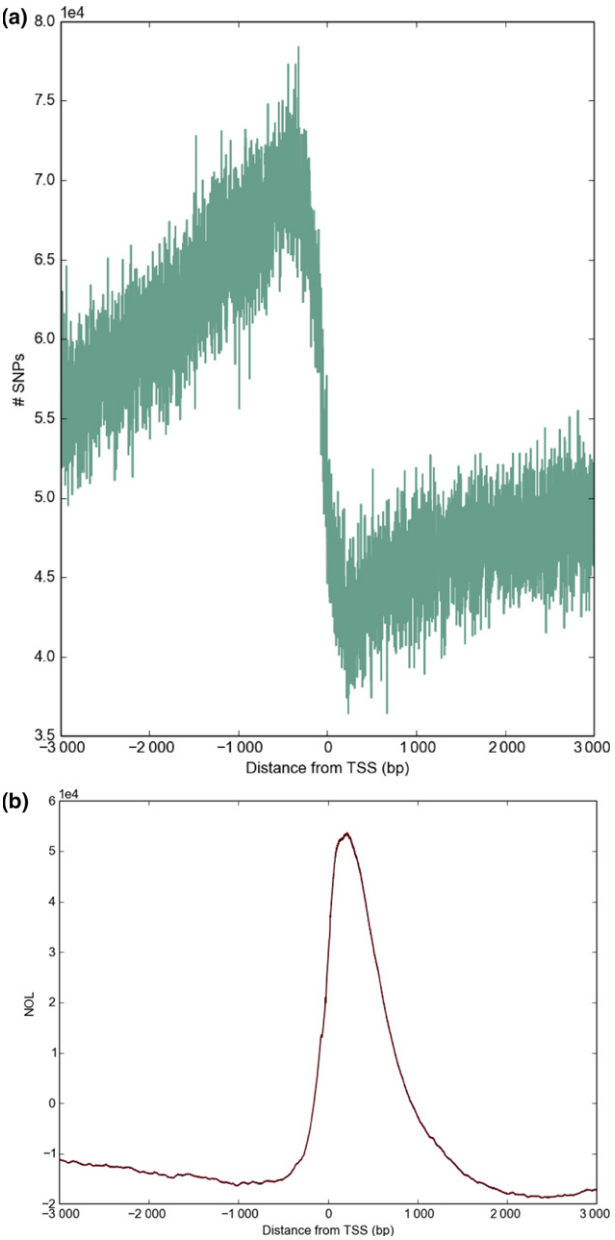


**Figure 4.** Genetic variation and nucleosome occupancy likelihoods around transcription start sites in the sorghum genome.
(a, b) Consistent with experimental observations in maize and *Arabidopsis*, nucleosome occupancy scores indicate a high likelihood of a nucleosome positioned immediately downstream of transcription start sites in sorghum. Unlike previous observations in medaka and human, strong evidence that nucleosomes were periodically arrayed at the TSS was not observed in either the accumulation of genetic variants nor nucleosome occupancy likelihoods.

(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 (Witte *et al.*, 2001; Gao *et al.*, 2016). 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 *Sorghum 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; Figures S2 and S4).

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.

### 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: (i) to sample the major plant organs (roots, leaves, stems, panicles) at different developmental stages (juvenile, vegetative, reproductive) to facilitate comprehensive annotation of genes in the sorghum genome; and (ii) 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 Table S2 and Data 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

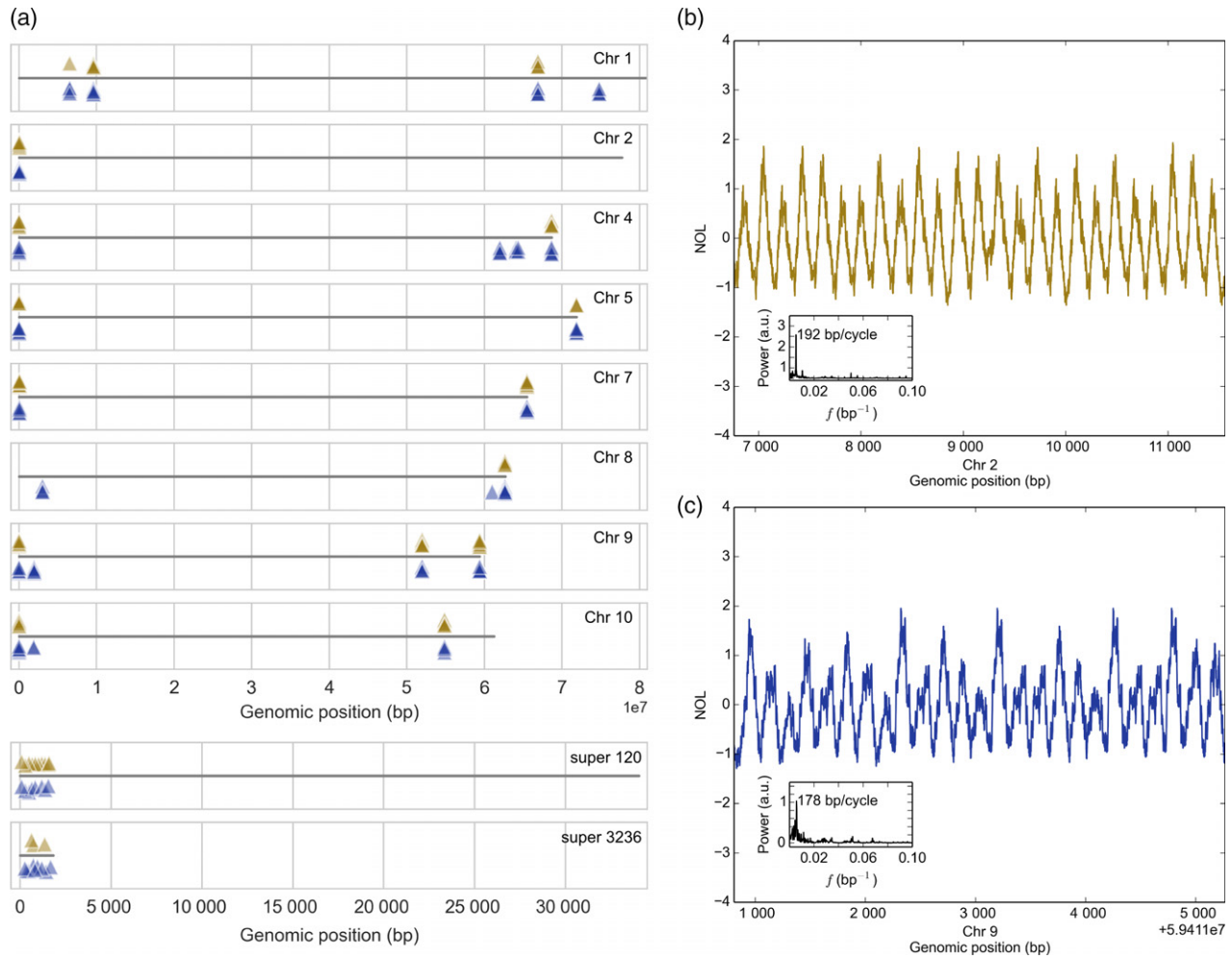


**Figure 5.** Subtelomeric periodicities in nucleosome occupancy likelihoods correspond to arrays of tandem repeats located near the end of most chromosome arms.
(a) Graphic representation of BLAST hits for the consensus sequence of STA1 and STA2 indicate that most chromosome arms contain subtelomeric 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.

biological replicates of the same sample, as well as correlated groups between samples from the same tissue (Figure S5). 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 of 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 the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) method 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).

To identify a set of genes with large variation in expression across the dataset, principal component analysis was performed using standardized expression values of 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

 

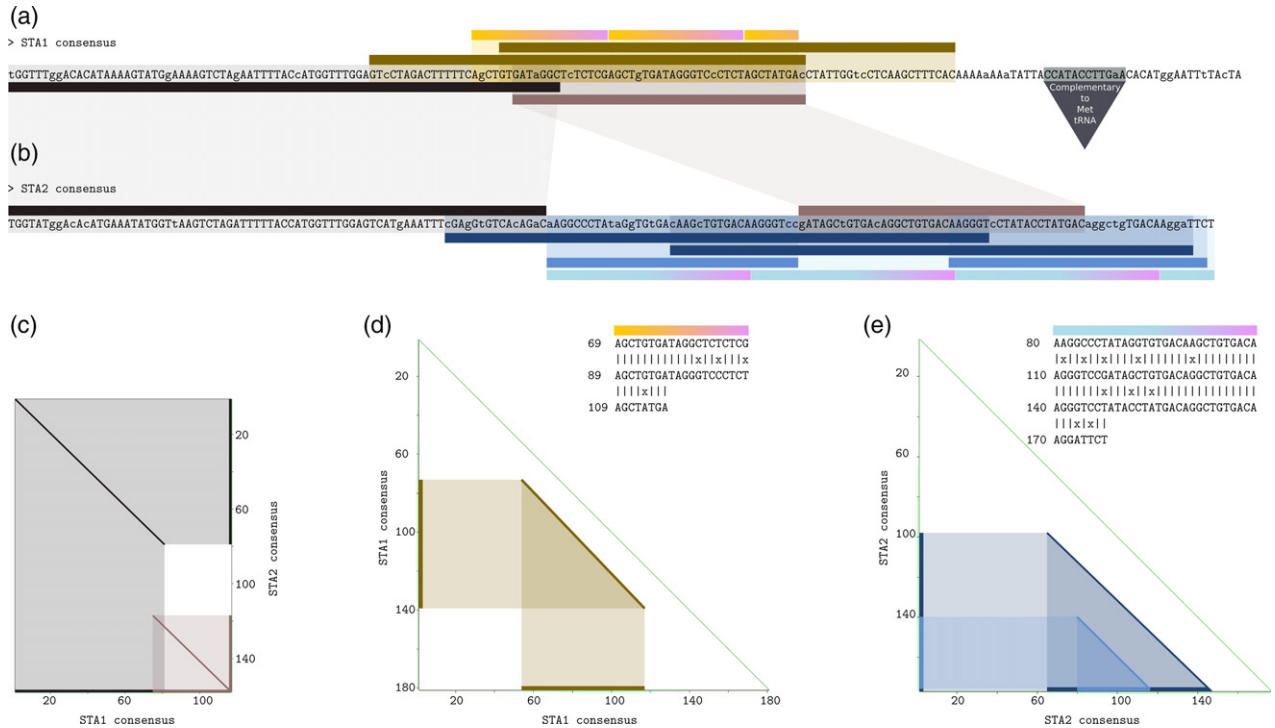


**Figure 6.** Relationships and structure of STA monomers.
Monomers of STA1 and STA2 share blocks of identity and each contain direct repeats.
(a, b) Consensus sequence of the STA1 and STA2 monomers annotated for sequence shared between monomers as well as direct repeats within each monomer.
(c–e) Annotations are indicated by colored bars and color coded according to their origin, either from (c) an alignment dot plot of STA1 and STA2, (d) a self-alignment dot plot and tandem direct repeat analysis of STA1, or (e) a self-alignment dot plot and tandem direct repeat analysis of STA2. STA1 also contains a putative primer binding site (PBS) complementary to a sorghum methionine tRNA, a feature characteristic of TRIM elements (Witte *et al.*, 2001; Chan and Lowe, 2009).

identified (Figure S6). 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), protein kinase activity (GO0004672), DNA-directed RNA polymerase activity (GO0003899) (Data S4). As kinase activity is associated with signal transduction and the other two terms may be explained by the developmental state of the tissue (e.g. high transcription and translation activity), the 47 genes responsible for protein kinase activity were investigated further (Table S3).

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

## DISCUSSION

The sorghum genome sequencing project was organized in 2005 because *S. bicolor* has a relatively simple genome compared to many other grasses, sorghum is a valuable genetic model for C4 grass research, and sorghum crops are important worldwide, especially as subsistence crops in the semi-arid tropics (sorghum Genomics Planning Workshop Participants, 2005). The sorghum genome sequence improved our understanding of sorghum genome organization, coding capacity, and aided analysis of grass genome diversification (Paterson *et al.*, 2009). The original reference genome sequence was based on approximately 8.5-fold depth paired-end Sanger sequence reads from genomic libraries with 100-fold variation in the size of inserts. Discrepancies in order that arose during assembly were resolved in part using information from pre-existing high resolution genetic and BAC-based physical maps of the sorghum genome (Klein *et al.*, 2000; Bowers *et al.*, 2003). The sum of the 201 largest sequenced scaffolds spanned 678.9 Mbp of which 625.7 Mbp of the sequence was assigned chromosomal locations. The size of the reference genome had been previously estimated by flow

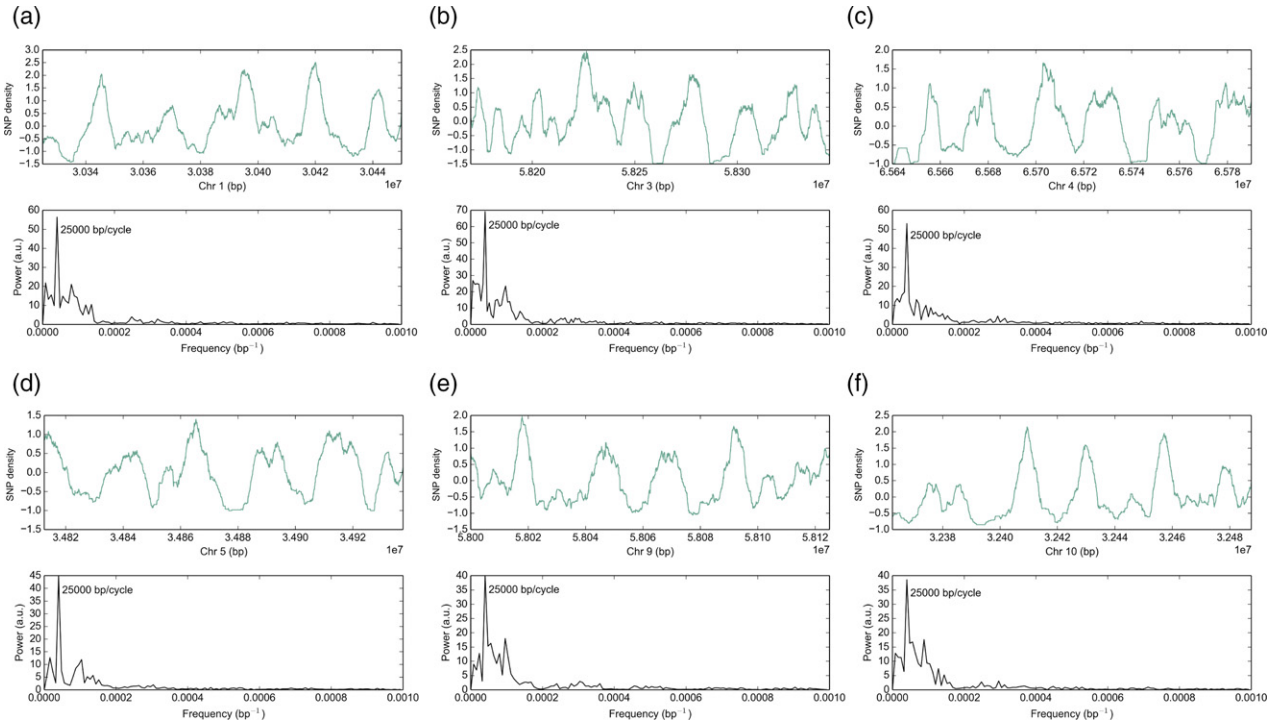


**Figure 7.** Periodicities in the accumulation of genetic variation in the sorghum genome.
(a–f) A genome-wide scan for periodic accumulation of SNPs identified multiple regions of the genome with a distinct period of 25 000 bp. The top plot of each panel shows the scaled SNP density relative to the 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.

cytometry to be 818 Mbp (Price *et al.*, 2005), indicating that reference genome v1 sequence comprising the 10 chromosomes accounted for approximately 76% of the total genome sequence. It was reported that 15 of the 20 chromosome termini contained telomeric repeats and that Cen38 sequences (Zwick *et al.*, 2000) were present in each chromosome, although these sequences were also found in many of the sequence scaffolds that could not be incorporated into the chromosomal sequences (Paterson *et al.*, 2009). Despite the need for further improvement, the resulting sorghum reference sequence has been of great value to the sorghum and grass research community, enabling comparative genomics (Paterson *et al.*, 2009), association studies (e.g. Morris *et al.*, 2013; Brenton *et al.*, 2016), the development of genotyping by sequencing methods for sorghum (Morishige *et al.*, 2013), analysis of sorghum diversity and variant distribution (Evans *et al.*, 2013; Mace *et al.*, 2013; McCormick *et al.*, 2015), genome methylation profiles (Olson *et al.*, 2014), and many other research activities.

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

Version 1.4 of the sorghum genome sequence annotation provided evidence for 27 604 annotated genes. Subsequent analysis of gene annotations that incorporated RNA-seq data indicated that a large number of genes were not annotated in v1.4 and that many of the annotations were incomplete (Olson *et al.*, 2014). Results from the current study based on the improved genome sequence, an updated gene model identification pipeline, and deep RNA-seq analysis of 47 tissues/developmental stages enabled

 

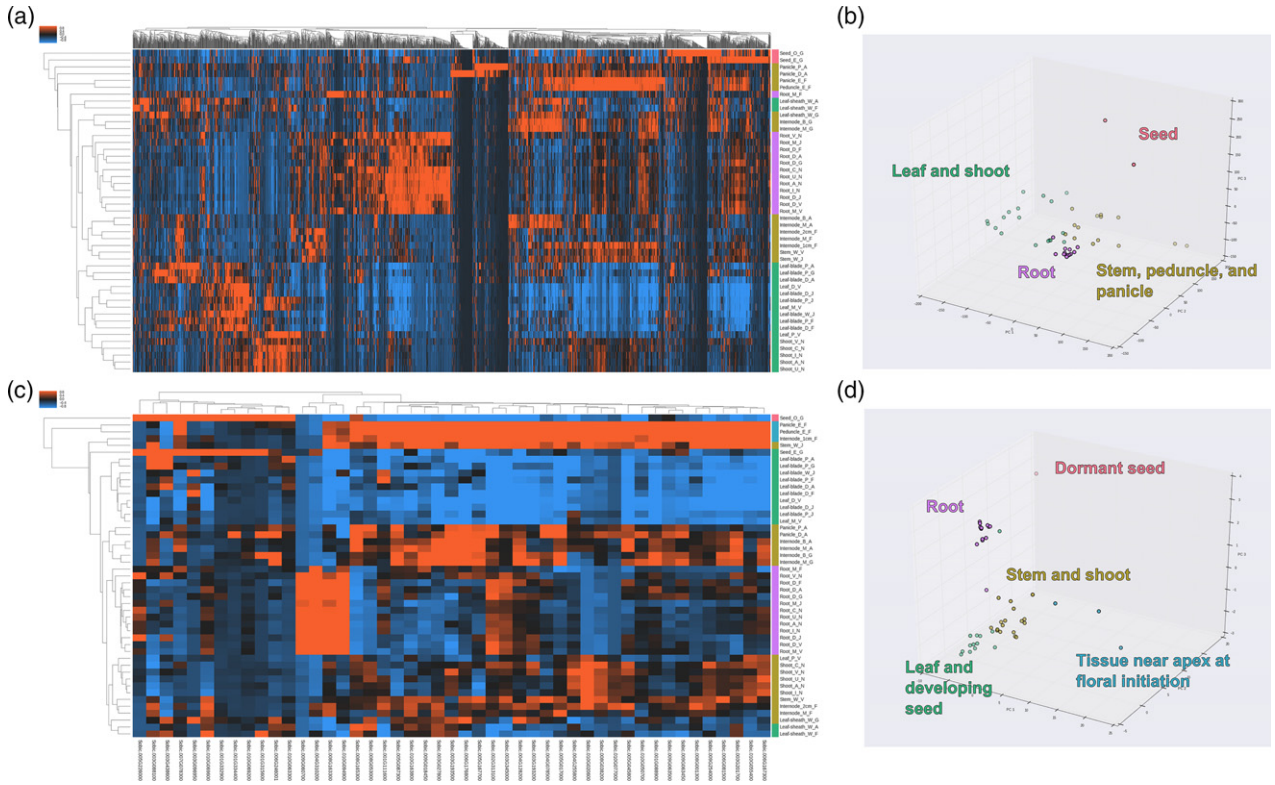


**Figure 8.** Clustering and ontological analyses indicate the expression of kinase genes are associated with tissue identity.
(a) Heat map and hierarchical clustering of atlas samples based on gene expression of all 34 211 sorghum genes; color bars on right correspond to *k*-means clusters in panel (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.

the annotation of 34 211 genes, a 24% increase relative to v1.4. Gene models required evidence of gene expression and/or homology to a known gene (protein) to be included in the gene annotation list; however, as would be expected, the evidence for a given gene model varies. Of the 27 604 genes annotated in v1.4, 22 483 genes completely overlap and align to a v3.1 gene annotation. Several thousand additional v1.4 genes align imperfectly to genes in v3.1 but those with 70% or greater correspondence were identified. Information on the correspondence of v1.4 to v3.1 gene annotations can be accessed from Phytozome (phytozome.jgi.doe.gov). In addition, 6 989 new gene annotations that had no corresponding gene annotation in v1.4 were added, increasing the total gene count in v3.1 to 34 211. The additional genes are enriched in several gene ontologies and KEGG pathways related to photosynthesis and oxidative phosphorylation. A preliminary analysis of the entire gene set showed that approximately 1 151 genes out of 34 211 were not expressed in the RNA atlas, although many of these genes were homologs of genes annotated in Arabidopsis and rice. This finding indicates that the RNA atlas needs further development so that it includes the expression of genes induced in response to biotic and abiotic stress, hormones, and specific cell types. In addition, 688 of the 34 211 gene annotations aligned to transposons or domains of proteins encoded by transposons in rice or Arabidopsis. These genes were included in the 34 211 because they were low copy number genes. Taken together, v3.1 now contains 34 211 annotated genes, a 24% increase over v1.4 annotations. Emphasis was placed on being inclusive to avoid false negatives; however, we hope that future studies by the sorghum and grass research communities including pan-genome analyses will provide an even better defined set of sorghum genes.

RNA-seq data improved the annotation of exons resulting in a significant increase in average gene size consistent with prior results based on a similar approach (Olson *et al.*, 2014). A comparison of 23 135 genes from v1.4 to the matching genes in v3.1 (100% overlap) revealed that 7 646 of the v3.1 models have longer CDS lengths with an

average increase in length of 400 bp (Figure S3). Increased gene coverage and improved gene annotation and sequence accuracy will aid comparative genomics studies as well as GWAS and map-based QTL to gene discovery projects. Such projects can result in false negatives/positives if the reference genome sequence used for analysis is not a well annotated high-quality sequence. In our own research, errors and misannotation of the v1.4 sequence caused identification of candidate gene alleles underlying QTL to be missed until direct sequencing was carried out on all genes in fine mapped intervals (Murphy *et al.*, 2011; Hilley *et al.*, 2016).

While v3.1 is a substantial improvement over v1.4, additional information is needed to fill in missing portions of the genome sequence and to improve gene annotation. As noted above, one end of five chromosomes lacked telomeric sequences, indicating that these chromosome sequences were not complete. Moreover, it is likely that the sequence of the pericentromeric repeat-rich regions of chromosomes was incomplete and possibly misordered in some regions. As 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 showed that substantial numbers of 'dispensable' genes were 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, 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 (Witte *et al.*, 2001; Gao *et al.*, 2016). The function of these subtelomeric repeats is unknown, although subtelomeric repeats have been shown to be involved in both bouquet formation and in facilitating the pairing of homologous chromosomes during meiosis (Sadaie *et al.*, 2003; Harper *et al.*, 2004). A complete analysis of these subtelomeric arrays will require additional long-read sequencing to fully characterize the size and location of these subtelomeric repeats and to determine if they are present in all of the sorghum chromosomes.

Comparison of whole-genome sequences from 52 diverse sorghum genotypes to the v3 reference genome sequence identified approximately 7.4 M SNPs and approximately 1.9 M indels. Large-scale signals in the accumulation of genetic variation were identified by signal processing techniques, and these may represent signatures left by higher order organization. For example, elevated variant frequency was associated with the wobble position in codons. Previous studies have documented elevated variant density in genes and repeat sequences of euchromatic regions relative to pericentromeric regions of sorghum chromosomes and significant variation in variant density within euchromatin when the genomes of different sorghum races were compared (Evans *et al.*, 2013). Genetic hitchhiking may be acting to reduce genetic variation in regions of low recombination near centromeres (Barton, 2000). In this study, variant distributions based on the analysis of 52 sorghum genomes were analyzed and found to contain large-scale variant distribution features that repeat every approximately 25 kbp. We had previously speculated that large-scale features like these could be generated by regional variation in recombination and repair, possibly due to higher order chromatin organization (Evans *et al.*, 2013). In addition, the ability of DNA repair machinery to access and correct mutations and selection pressures generated by functional properties of the genome such as gene coding sequences across the gene rich distal arms of sorghum chromosomes in which rates of recombination are high could be influencing the accumulation of variants (Zheng *et al.*, 2011; Evans *et al.*, 2013; Mace *et al.*, 2013; Makova and Hardison, 2015). Follow on studies could leverage wavelet transforms in addition to the DFT to

resolve problems associated with non-stationary signals, as these genomic signals are likely non-stationary in nature. Moreover, wavelet transform coefficients can be used to correlate multiple features such as recombination and genetic variation (Spencer *et al.*, 2006). The results from digital signal processing approaches used to examine the sorghum genome indicate that additional experimentation to annotate sorghum chromatin, including higher order features like chromatin interactions and nuclear lamina binding sites, will be useful to better understand factors shaping the landscape of the sorghum genome; such experiments will determine if these signals are generated by biological processes.

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

## EXPERIMENTAL PROCEDURES

### Genome assembly and improvement

In total, 320 regions of the version 1 sorghum reference genome assembly (Paterson *et al.*, 2009) that contained a gene density greater than two 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 between BTx623 and IS3620C was used to both improve the quality of the assembly and to 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 eight breaks were identified in the v1.0 release chromosomes, and an additional seven previously unmapped scaffolds were integrated into the assembly in the appropriate location (Data S1). A 1.08 Mb region of the v1.0 chromosome 6 was moved to chromosome 7. Fifteen 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 110× of 2 × 250 (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 150 bp of one another were not corrected. Additional information regarding methods of assembly and finishing are contained in Data S1.

### Sample preparation and sequencing for transcriptome atlas and whole-genome resequencing

The reference line BTx623 was grown under 14 h day lengths in greenhouse conditions in topsoil, equivalent to native field soil from Brazos County, TX, to generate tissue for two separate experiments: (i) a tissue by developmental stage time course; and (ii) a nitrogen source study. For the tissue by developmental stage time course, plants were harvested at the juvenile stage (8 DAE),

the vegetative stage (24 DAE), at floral initiation (44 DAE), at anthesis (65 DAE), and at grain maturity (96 DAE). Leaf, root, stem and reproductive structures were flash frozen in liquid nitrogen. For each tissue by stage combination, three biological replicates (i.e. three plants representing a single condition) were harvested with the exception of the juvenile stage, for which a replicate was represented by five plants instead of one to compensate for lower tissue abundance. For the nitrogen source study, plants grown under differing nitrogen source regimes were harvested at 30 DAE, and shoots and roots were flash frozen. For each tissue by condition, three biological replicates were obtained. Additional details regarding harvested samples can be found in Table S2 and Data S1 and S3.

Tissue was ground under liquid nitrogen and RNA was extracted using a TRIzol-reagent based extraction. Tissues with high levels of starch used a modified TRIzol-reagent protocol (Li and Trick, 2005). Plate-based RNA sample prep was performed on the PerkinElmer Sciclone NGS robotic liquid handling system using Illumina's TruSeq Stranded mRNA HT sample prep kit utilizing poly-A selection of mRNA following the protocol outlined by Illumina in their user guide: http://support.illumina.com/sequencing/sequencing_kits/truseq_stranded_mrna_ht_sample_prep_kit.html, and with the following conditions: total RNA starting material was 1 μg per sample and eight cycles of PCR was used for library amplification. The prepared libraries were then quantified by qPCR using the Kapa SYBR Fast Illumina Library Quantification Kit (Kapa Biosystems) and run on a Roche LightCycler 480 real-time PCR instrument. The quantified libraries were then prepared for sequencing on the Illumina HiSeq sequencing platform utilizing a TruSeq paired-end cluster kit, v4, and Illumina's cBot instrument to generate a clustered flow cell for sequencing. Sequencing of the flow cell was performed on the Illumina HiSeq 2 500 sequencer using HiSeq TruSeq SBS sequencing kits, v4, following a 2 × 150 indexed run recipe. Sequencing generated roughly 3.3 billion pairs of sorghum paired-end read data.

Nine additional sorghum lines (100 M, 80 M, BTx623, BTx642, Hegari, IS3620C, SC170-6-17, Standard Broomcorn, and Tx7000) were resequenced to supplement the 47 lines already available (Zheng *et al.*, 2011; Mace *et al.*, 2013). Seeds were soaked in 20% bleach for 20 min and washed extensively in distilled water for 1 h. Seeds were germinated on water saturated germination paper in a growth chamber (14 h light; 30°C/10 h dark; 24°C). Genomic DNA was isolated from 8-day-old root tissue using a FastPrep DNA Extraction kit and FastPrep24 Instrument (MP Biomedicals LLC, Solon, OH, USA), according to the manufacturer's specifications. DNA template (350-bp average insert size) was prepared using a TruSeq® DNA PCR-Free LT Kit, according to the manufacturer's directions. Paired-end sequencing (125 × 125 bases) was performed on an Illumina HiSeq 2500.

### Transcriptome annotation

The RNA-seq reads were aligned to the updated reference assembly using GSNAP and assembled into 127 415 RNA-seq transcripts with the PERTRAN pipeline (Shu *et al.*, 2013). These transcripts were combined with 209 835 ESTs (obtained from GenBank) to generate 111 994 transcript assemblies using PASA. Loci were determined by transcript assembly alignments and/or EXONERATE alignments of proteins from *Arabidopsis thaliana*, rice, maize or grape genomes. Gene models were predicted by homology-based predictors, mainly FGENESH+, FGENESH_EST, and GenomeScan. The best scored predictions for each locus were selected using multiple positive factors including EST and protein support, and one negative factor: overlap with repeats. The selected gene predictions were improved by PASA by adding UTRs, splicing correction, and adding alternative transcripts. Finally, a homology analysis was performed on the PASA-improved models relative to the proteomes of *A. thaliana*, rice, maize and grape to identify high-quality gene models and remove models with extensive transposable element domains.

### Additional feature annotation, feature coverage, and periodicity analyses

Additional features were annotated in the sorghum genome, including repetitive sequence, genetic variants, and NOLs. Repetitive sequence, including transposons and SSRs, were annotated using both a *de novo* annotation and an annotation with existing libraries with REPET v2.5; existing repetitive element libraries included the TIGR Plant Repeat Database and RepBase (Ouyang and Buell, 2004; Quesneville *et al.*, 2005; Flutre *et al.*, 2011; Bao *et al.*, 2015). Genetic variants were called from sequence data for 56 sorghum resequenced sorghum samples (Data S5). Processing of sequence reads to variant calls, including alignment to the Sbi3 reference genome, base recalibration, indel realignment, joint genotyping, and variant quality score recalibration were performed using BWA v0.7.12 and GATK v3.3 and following the informed pipeline of the RIG workflow (Li and Durbin, 2009; McKenna *et al.*, 2010; DePristo *et al.*, 2011; Auwera *et al.*, 2013; McCormick *et al.*, 2015). Four of the 56 lines were excluded from subsequent analyses due to their evolutionary divergence (i.e. two *S. propinquum* genotypes and two subsp. *verticilliflorum* genotypes were excluded). For examining variant accumulation at TSSs 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 NOLs, the SVM 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.

### Characterization of STA1 and STA2

Sequences corresponding to STA1 and STA2 were identified initially by examining sequences underlying periodic NOLs. The STA1 and STA2 monomers were defined by finding the minimum complete repeat (approximately 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 nine different monomers representing a trio of tandem repeats from three different arrays on three different chromosome arms (Figures S2 and S4) 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).

 

### 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).

### Data access

The sorghum reference genome sequence and annotation are available from Phytozome (phytozome.jgi.doe.gov). The sequence has also been deposited in GenBank under accession number ABXC00000000. The chloroplast sequence is available as GenBank Accession EF115542, and the mitochondrion sequence is available as NCBI Ref Seq accession NC_008360. 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 Data S5; the nine lines sequenced as part of this work are associated with BioProject PRJNA374837. RNA-seq reads for the atlas are available from the NCBI SRA under accessions SRA558272, SRA558514 and SRA558539. Variant calls and RNA-seq data are also hosted on the Phytozome JBrowse genome browser for browsing and download (phytozome.jgi.doe.gov).

## ACKNOWLEDGEMENTS

The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US 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 Science BER DE-FC02-07ER64494), and the US Department of Energy grants no. DE-AR0000596 and DE-SC0012629. The authors have no conflicts of interest to declare.

## AUTHORS' CONTRIBUTIONS

RFM and SKT performed downstream analyses (e.g. expression clustering, coverage analyses, periodicity analyses), transposon annotation, and linkage analyses. AS performed RNA-seq QC, read mapping and expression analyses. SS performed gene annotation (gene set version 3.1). JJ, DS, and JG performed genome assembly and finishing (genome version 3.0). MK and MA performed sorghum transcriptome atlas sequencing. RFM, SKT, BW, BM, and AM prepared transcriptome atlas samples. DM performed resequencing of selected sorghum lines. JG, JS, and JM conceived and provided project management. RFM, SKT, and JM wrote the manuscript. All authors reviewed and approved of the manuscript.

## SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

**Figure S1.** Coverage maps of features for chromosomes 1–6.

**Figure S2.** Coverage maps of features for chromosomes 7–10.

**Figure S3.** Length of coding sequence increases between gene annotations 1.4 and 3.1.

**Figure S4.** Multiple sequence alignment of nine STA1 and nine STA2 sequences corresponding to three monomers from each of three arrays on different chromosome arms.

**Figure S5.** Gene expression correlations for all samples and replicates.

**Figure S6.** Loadings for first three principal components of each gene.

**Table S1.** Presence and absence of telomere repeat and STA repeats in the sorghum genome.

**Table S2.** Tissue samples present in the sorghum transcriptome atlas.

**Table S3.** Genes annotated as having protein kinase activity responsible for ontological enrichment in the 2500 genes with large principal component loadings.

**Data S1.** Supplemental methods.

**Data S2.** STA sequence.

**Data S3.** Transcriptome Atlas ID mapping.

**Data S4.** Ontological enrichments.

**Data S5.** Whole-genome sequence accessions.

## 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. *Nat. Commun.* **7**, 11706.

**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 Protoc. Bioinformatics*, **43**, 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. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **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. *J. Exp. Bot.* **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. *Nat. Biotechnol.* **30**, 555–561.

**Bonev, B. and Cavalli, G.** (2016) Organization and function of the 3D genome. *Nat. Rev. Genet.* **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. *J. Plant. Regist.* **5**, 141–145.

**Chan, P.P. and Lowe, T.M.** (2009) GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucl. Acids Res.* **37**, D93–D97.

**Corpet, F.** (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* **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. *Nat. Genet.* **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 Genom.* **12**, 514.

**ENCODE Project Consortium** (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

**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 Physiol.* **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 Biol.* **17**, 7.

**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. *J. Cell Sci.* **117**, 4025–4032.

**Higasa, K. and Hayashi, K.** (2006) Periodicity of SNP distribution around transcription start sites. *BMC Genom.* **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. *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. *Nat. Rev. Genet.* **12**, 756–766.

**International Barley Genome Sequencing Consortium** (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature*, **491**, 711–716.

**International Human Genome Sequencing Consortium** (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

**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 Physiol.* **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. *et al.*** (2000) A high-throughput AFLP-based method for constructing integrated genetic and physical maps: progress toward a sorghum genome map. *Genome Res.* **10**, 789–807.

**Kohorn, B.D.** (2015) The state of cell wall pectin monitored by wall associated kinases: a model. *Plant. Signal. Behav.* **10**, e1035854.

**Kolpakov, R., Bana, G. and Kucherov, G.** (2003) mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res.* **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 Res.* **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. *Nat. Commun.* **4**, 2320.

**Makova, K.D. and Hardison, R.C.** (2015) The effects of chromatin organization on variation in mutation rates in the genome. *Nat. Rev. Genet.* **16**, 213–223.

**McCormick, R.F., Truong, S.K. and Mullet, J.E.** (2015) RIG: recalibration and interrelation of genomic sequence data with the GATK. *G3*, **5**, 655–665.

**McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S. and Daly, M.** (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303.

**McKinley, B., Rooney, W., Wilkerson, C. and Mullet, J.** (2016) Dynamics of biomass partitioning, stem gene expression, cell wall biosynthesis, and sucrose accumulation during development of *Sorghum bicolor*. *Plant J.* **88**, 662–680.

**Mehrotra, S. and Goyal, V.** (2014) Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. *Genomics Proteomics Bioinformatics*, **12**, 164–171.

**Mickelbart, M.V., Hasegawa, P.M. and Bailey-Serres, J.** (2015) Genetic mechanisms of abiotic stress tolerance that translate to crop yield stability. *Nat. Rev. Genet.* **16**, 237–251.

**Mondal, S., Rutkoski, J.E., Velu, G., Singh, P.K., Crespo-Herrera, L.A., Guzman, C.G., Bhavani, S., Lan, C., He, X. and Singh, R.P.** (2016) Harnessing diversity in wheat to enhance grain yield, climate resilience, disease and insect pest resistance and nutrition through conventional and modern breeding approaches. *Front. Plant Sci.* **7**, 991.

**Morishige, D.T., Klein, P.E., Hilley, J.L., Sahraeian, S.M.E., Sharma, A. and Mullet, J.E.** (2013) Digital genotyping of sorghum – a diverse plant species with a large repeat-rich genome. *BMC Genom.* **14**, 448.

**Morris, G.P., Ramu, P., Deshpande, S.P., Hash, C.T., Shah, T., Upadhyaya, H.D., Riera-Lizarazu, O., Brown, P.J., Acharya, C.B. and Mitchell, S.E.** (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc. Natl Acad. Sci. USA*, **110**, 453–458.

**Mullet, J., Morishige, D., McCormick, R., Truong, S., Hilley, J., McKinley, B., Anderson, R., Olson, S.N. and Rooney, W.** (2014) Energy sorghum-a genetic model for the design of C-4 grass bioenergy crops. *J. Exp. Bot.* **65**, 3479–3489.

**Murphy, R.L., Klein, R.R., Morishige, D.T., Brady, J.A., Rooney, W.L., Miller, F.R., Dugas, D.V., Klein, P.E. and Mullet, J.E.** (2011) Coincident light and clock regulation of pseudoresponse regulator protein 37 (PRR37) controls photoperiodic flowering in sorghum. *Proc. Natl Acad. Sci. USA*, **108**, 16469–16474.

**Nakamichi, N., Murakami-Kojima, M., Sato, E., Kishi, Y., Yamashino, T. and Mizuno, T.** (2002) Compilation and characterization of a novel WNK family of protein kinases in *Arabiodpsis thaliana* with reference to circadian rhythms. *Biosci. Biotechnol. Biochem.* **66**, 2429–2436.

**Noé, L. and Kucherov, G.** (2005) YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res.* **33**, W540–W543.

**Olson, A., Klein, R.R., Dugas, D.V., Lu, Z., Regulski, M., Klein, P.E. and Ware, D.** (2014) Expanding and vetting sorghum bicolor gene annotations through transcriptome and methylome sequencing. *Plant Genome*, **7**. https://doi.org/10.3835/plantgenome2013.08.0025

**Ort, D.R., Merchant, S.S., Alric, J. *et al.*** (2015) Redesigning photosynthesis to sustainably meet global food and bioenergy demand. *Proc. Natl Acad. Sci. USA*, **112**, 8529–8536.

**Ouyang, S. and Buell, C.R.** (2004) The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res.* **32**, D360–D363.

**Padeken, J., Zeller, P. and Gasser, S.M.** (2015) Repeat DNA in genome organization and stability. *Curr. Opin. Genet. Dev.* **31**, 12–19.

**Park, S.-Y., Peterson, F.C., Mosquna, A., Yao, J., Volkman, B.F. and Cutler, S.R.** (2015) Agrochemical control of plant water use using engineered abscisic acid receptors. *Nature*, **520**, 545–548.

**Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T. and Poliakov, A.** (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, **457**, 551–556.

 

**Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V.** (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830.

**Price, H.J., Dillon, S.L., Hodnett, G., Rooney, W.L., Ross, L. and Johnston, J.S.** (2005) Genome evolution in the genus sorghum (Poaceae). *Ann. Bot.* **95**, 219–227.

**Quesneville, H., Bergman, C.M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M. and Anxolabehere, D.** (2005) Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput. Biol.* **1**, e22.

**Rosenbloom, K.R., Sloan, C.A., Malladi, V.S., Dreszer, T.R., Learned, K., Kirkup, V.M., Wong, M.C., Maddren, M., Fang, R. and Heitner, S.G.** (2013) ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res.* **41**, D56–D63.

**Sadaie, M., Naito, T. and Ishikawa, F.** (2003) Stable inheritance of telomere chromatin structure and function in the absence of telomeric repeats. *Genes Dev.* **17**, 2271–2282.

**Sanborn, A.L., Rao, S.S., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A. and Li, J.** (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl Acad. Sci. USA*, **112**, E6456–E6465.

**Sasaki, S., Mello, C.C., Shimada, A., Nakatani, Y., Hashimoto, S.-I., Ogawa, M., Matsushima, K., Gu, S.G., Kasahara, M. and Ahsan, B.** (2009) Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. *Science*, **323**, 401–404.

**Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L. and Graves, T.A.** (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.

**Sekhon, R.S., Lin, H., Childs, K.L., Hansey, C.N., Buell, C.R., de Leon, N. and Kaeppler, S.M.** (2011) Genome-wide atlas of transcription during maize development. *Plant J.* **66**, 553–563.

**Sekhon, R.S., Briskine, R., Hirsch, C.N., Myers, C.L., Springer, N.M., Buell, C.R., de Leon, N. and Kaeppler, S.M.** (2013) Maize gene atlas developed by RNA sequencing and comparative evaluation of transcriptomes based on RNA sequencing and microarrays. *PLoS ONE*, **8**, e61005.

**Shakoor, N., Nair, R., Crasta, O., Morris, G., Feltus, A. and Kresovich, S.** (2014) A *Sorghum bicolor* expression atlas reveals dynamic genotype-specific expression profiles for vegetative tissues of grain, sweet and bioenergy sorghums. *BMC Plant Biol.* **14**, 1.

**Shpak, E.D., Berthiaume, C.T., Hill, E.J. and Torii, K.U.** (2004) Synergistic interaction of three ERECTA-family receptor-like kinases controls Arabidopsis organ growth and flower development by promoting cell proliferation. *Development*, **131**, 1491–1501.

**Shu, S., Goodstein, D.M. and Rokhsar, D.** (2013) PERTRAN: genome-guided RNA-seq Read Assembler. *Methods*, **40**, 50.

**Smékalová, V., Luptovčiak, I., Komis, G., Šamajová, O., Ovečka, M., Doskočilová, A., Takáč, T., Vadovič, P., Novák, O. and Pechan, T.** (2014) Involvement of YODA and mitogen activated protein kinase 6 in Arabidopsis post-embryogenic root development through auxin up-regulation and cell division plane orientation. *New Phytol.* **203**, 1175–1193.

**Sorghum Genomics Planning Workshop Participants** (2005) Toward sequencing the sorghum genome. A US National Science Foundation-sponsored workshop report. *Plant Physiol.* **138**, 1898–1902.

**Spencer, C.C., Deloukas, P., Hunt, S., Mullikin, J., Myers, S., Silverman, B., Donnelly, P., Bentley, D. and McVean, G.** (2006) The influence of recombination on human genetic diversity. *PLoS Genet.* **2**, e148.

**Stahl, Y., Grabowski, S., Bleckmann, A., Kühnemuth, R., Weidtkamp-Peters, S., Pinto, K.G., Kirschner, G.K., Schmid, J.B., Wink, R.H. and Hülsewede, A.** (2013) Moderation of Arabidopsis root stemness by CLAVATA1 and ARABIDOPSIS CRINKLY4 receptor kinase complexes. *Curr. Biol.* **23**, 362–371.

**Tang, H., Klopfenstein, D., Pederson, B., Flick, P., Sato, K., Ramirez, F., Yunes, J. and Mungall, C.** (2015) GOATOOLS: tools for gene ontology. *Zendo*. https://doi.org/10.5281/zenodo.31628

**Technow, F., Messina, C.D., Totir, L.R. and Cooper, M.** (2015) Integrating crop growth models with whole genome prediction through approximate Bayesian computation. *PLoS ONE*, **10**, e130588.

**Tolstorukov, M.Y., Volfovsky, N., Stephens, R.M. and Park, P.J.** (2011) Impact of chromatin structure on sequence variability in the human genome. *Nat. Struct. Mol. Biol.* **18**, 510–515.

**Truong, S.K., McCormick, R.F., Morishige, D.T. and Mullet, J.E.** (2014) Resolution of genetic map expansion caused by excess heterozygosity in plant recombinant inbred populations. *G3*, **4**, 1963–1969.

**Uddin, M.N., Dunoyer, P., Schott, G., Akhter, S., Shi, C., Lucas, W.J., Voinnet, O. and Kim, J.-Y.** (2014) The protein kinase TOUSLED facilitates RNAi in Arabidopsis. *Nucleic Acids Res.* **42**, 7971–7980.

**Van Buren, R., Bryant, D., Edger, P.P., Tang, H., Burgess, D., Challabathula, D., Spittle, K., Hall, R., Gu, J. and Lyons, E.** (2015) Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature*, **527**, 508–511.

**Vogel, J.P., Garvin, D.F., Mockler, T.C., Schmutz, J., Rokhsar, D., Bevan, M.W., Barry, K., Lucas, S., Harmon-Smith, M. and Lail, K.** (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, **463**, 763–768.

**Voytas, D.F.** (2013) Plant genome engineering with sequence-specific nucleases. *Plant Biol.* **64**, 327.

**Wang, Y., Liu, K., Liao, H., Zhuang, C., Ma, H. and Yan, X.** (2008) The plant WNK gene family and regulation of flowering time in Arabidopsis. *Plant Biol.* **10**, 548–562.

**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 expression atlas covering the entire life cycle of rice. *Plant J.* **61**, 752–766.

**Witte, C.-P., Le, Q.H., Bureau, T. and Kumar, A.** (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proc. Natl Acad. Sci. USA*, **98**, 13778–13783.

**Zhang, G., Liu, X., Quan, Z., Cheng, S., Xu, X., Pan, S., Xie, M., Zeng, P., Yue, Z. and Wang, W.** (2012) Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nat. Biotechnol.* **30**, 549–554.

**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, C.-M.** (2011) Genome-wide patterns of genetic variation in sweet and grain sorghum (*Sorghum bicolor*). *Genome Biol.* **12**, 1.

**Zwick, M., Islam-Faridi, M., Zhang, H., Hodnett, G., Gomez, M., Kim, J., Price, H. and Stelly, D.** (2000) Distribution and sequence analysis of the centromere-associated repetitive element CEN38 of *Sorghum bicolor* (Poaceae). *Am. J. Bot.* **87**, 1757–1764.