Expanding and Vetting *Sorghum bicolor* Gene Annotations through Transcriptome and Methylome Sequencing

Andrew Olson, Robert R. Klein, Diana V. Dugas, Zhenyuan Lu, Michael Regulski, Patricia E. Klein, and Doreen Ware*

Abstract

With the emergence and subsequent advancement of next-generation sequence technology, detailed structural and functional characterization of genomes is readily attainable. Here, we have sampled the *Sorghum bicolor* methylome by shallow sequencing of HSO$_3^-$ (bisulfite)-treated DNA and have used these data to identify methylation patterns associated with high confidence gene models. We trained a classifier to predict functional gene models based on expression levels, methylation profiles, and sequence conservation. We have expanded the transcriptome atlas by sequencing RNA from meristematic tissues, florets, and embryos, and utilized this information to develop a more complete annotation of the *sorghum* transcriptome. Our gene annotations modify 60% of Sbi1.4 (version 1.4 of sorghum gene annotations) gene models. The updated models most often have extended untranslated region (UTR) annotations (18,105), but some show longer protein coding regions (5096) or previously unannotated alternative transcripts (6493). A phylogenetic analysis suggests that 800 genes are missing from annotation Sbi1.4 and 400 gene models are split. The new annotations resolve 50% of split gene models and include 30% of conserved genes missing from the Sbi1.4 annotation. Using our classifier, we identified a large set of 34,276 novel potentially functional transcribed regions. These transcribed regions include protein coding genes, non-coding RNAs, and other classes of gene products.

*SORGHUM* ([*Sorghum bicolor* (L.) Moench] is a C4 grass native to Africa, and its tolerance to drought and high temperature allows sorghum to thrive in the arid regions of Africa, Australia, Asia, and the Americas. While sorghum is primarily grown for grain and forage, sweet and high-biomass sorghums have recently emerged as dedicated bioenergy crops (Rooney et al., 2007). Sorghum has been used as a model C4 grass species owing to a relatively small genome (730 Mb) (Price et al., 2005), excellent genetic and germplasm resources (Dillon et al., 2007), and an evolutionary relationship to important crop species including maize, rice, and sugarcane (Devos and Gale, 2000; Paterson et al., 2000; Paterson et al., 2009b). The expansion of the sorghum genome relative to rice is largely pericentromeric-localized heterochromatin, and alignment of rice and sorghum genomes reveal similar quantities of euchromatin with largely collinear gene order (Kim et al., 2005).

In 2009, a team of international collaborators reported the sequence and annotation of the sorghum genome, releasing the assembly to the public (Paterson et al., 2009a). Gene annotation was performed using a combination of
homology-based and ab initio methods with expressed sequences from 15 monocots, including sorghum, maize, sugarcane, and rice. To aid in sorghum gene identification, protein information was also derived from *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, and rice. The 34,496 gene models resulting from the gene annotation were broken into bona fide high confidence (27,640), low confidence (5197), pseudogene (727), and transposon (932) models (Paterson et al., 2009a). Low confidence models differed from their high confidence counterparts in a lack of expressed sequence tag (EST) support, shorter gene length, fewer exons, or lack of homology. At present, the annotated sorghum genome of genotype BTx623 is the only publicly available whole-genome sequence for sorghum and has become the reference genomic platform of the worldwide sorghum community. Hence, the annotation accuracy of this model genome is critical for future functional and structural genome investigations in sorghum and related grass species.

As next-generation RNA sequencing (RNAseq) technology advances and costs decrease, gene expression compendia (atlases) are being assembled to provide a more detailed understanding of genomic biology and an even greater improvement in genome annotation. To this end, plant transcriptome atlases have been initiated for a number of species, in which various environmental perturbations (biotic or abiotic stress), developmental time courses, genotypes, and specific tissues or cell types are examined (Benedito et al., 2008; Gan et al., 2011; He et al., 2009; Jiao et al., 2009; Libault et al., 2010; Lu et al., 2010; Schreiber et al., 2009; Sekhon et al., 2011; Severin et al., 2010; Wang et al., 2010; Weber et al., 2007; Zhang et al., 2010). In sorghum, we previously utilized RNAseq technology to characterize the sorghum transcriptome and to examine differential gene expression in response to exogenous abscisic acid (ABA) and osmotic stress (Dugas et al., 2011). RNAseq has also been used to refine the annotation of plant genomes including *Arabidopsis* (Gan et al., 2011; Lamesch et al., 2012), rice (Lu et al., 2010; Zhang et al., 2010), cucumber (Li et al., 2011), date palm (Zhang et al., 2012), *Jatropha curcas* L. (Jiang et al., 2012), eastern cottonwood (*Populus deltoids*) (Yang et al., 2011), *Cannabis sativa* (van Bakel et al., 2011), carrot (Iorizzo et al., 2011), *Eucalyptus* (Mizrachi et al., 2010) and rubber tree (*Hevea brasiliensis* Muell. Arg) (Li et al., 2012a), among others. Through high-throughput transcriptome sequencing efforts, researchers have discovered new genes, transcriptionally-active pseudogenes, and noncoding RNAs; however, the greatest impact might be the structural refinement of existing gene models through refined 5’ and 3’ UTRs, and detection of alternative splicing events (Li et al., 2011; Lu et al., 2010; Marquez et al., 2012; Weber et al., 2007; Zhang et al., 2010). For example, alternative splicing analysis in *Arabidopsis* revealed increased numbers of splicing events and increased instances of GC-AG and AT-AC splice sites (Marquez et al., 2012; Syed et al., 2012). Transcriptome analyses of multiple reference genomes of *Arabidopsis* revealed novel genes, including 60 that have hits to the protein database (Weber et al., 2007). These improvements were incorporated into the latest TAIR release of the genome annotation (Lamesch et al., 2012). In rice, an initial transcriptome atlas that was based on eight plant organs revealed thousands of novel genes, alternative splicing events, and improved UTR annotation (Zhang et al., 2010). However, there are inherent tradeoffs and limitations imposed by short sequence reads, which are not inherent to conventional complementary DNA (cDNA) libraries. These limitations, which include a potential lack of strand specificity, intron read-through, and the difficulty in interpretation of complex short-sequence assemblies, often mandate that RNAseq data be complemented with protein, cDNA, or EST information or longer, strand-specific RNAseq sequence technology.

The advances of next-generation sequence technology have also facilitated an examination of epigenetic modifications of plant genomes. Methylation of cytosines and specific histone residues are two of the most heritable, but reversible, epigenetic modifications that occur (Baroux et al., 2002; Finnegan et al., 2000; Lippman et al., 2004; Schmitz and Ecker, 2012; Su et al., 2011). The method of choice for detection of DNA methylation is NaHSO$_3$ (sodium bisulfite) conversion of unmethylated cytosines to uracils followed by next-generation genome-wide sequencing (Bibikova and Fan, 2010; Clark et al., 1994; Cokus et al., 2008; Frommer et al., 1992; Harris et al., 2010; Krueger et al., 2012; Lister et al., 2008; Regulski et al., 2013; Reinders and Paszkowski, 2010; Zilberman et al., 2007). In higher plants, DNA methylation occurs at cytosines in a sequence context that is either symmetric (CpG [nucleotide pair; cytosine, guanine] and CHG [nucleotide triple; cytosine, not guanine, guanine], where H = A, T, or C) or asymmetric (CHH [nucleotide triple; cytosine, not guanine, guanine], where H = A, T, or C). Most methylated cytosines are found in repetitive and transposable elements, and conversely, hypomethylated regions may be actively transcribed or targeted by regulatory factors (Tran et al., 2005). However, gene coding and promoter regions can also be methylated, although the trans-generational maintenance of CpG methylation is more stable in transposable elements versus protein-coding genes (Becker et al., 2011; Schmitz and Ecker, 2012). Regulation of gene expression by DNA methylation is crucial for coordinating cellular developmental programs in many organisms, and DNA methylation in response to environmental conditions represents a potentially robust mechanism to regulate plant gene expression networks (Bilichak et al., 2012; Downen et al., 2012; Karan et al., 2012; Song et al., 2012; Tricker et al., 2012; Wang et al., 2011). In *Arabidopsis*, heritable epigenetic polymorphisms have been reported (Becker et al., 2011; Schmitz and Ecker, 2012), and major efforts are underway to sequence the DNA methylomes of other grasses and identify natural epigenetic variants (Eichten et al., 2011; Li et al., 2012b; Regulski et al., 2013; Springer, 2012). While the extent to which DNA epigenetic
variation contributes to the plant phenotype is unclear, characterizing the epigenome at a single-base resolution will likely be as critical as genotyping single nucleotide polymorphisms (SNPs) in understanding the complex interplay of the plant genome and the environment.

With an aim to develop methods for more complete functional and structural annotations of the sorghum genome, we have scanned the sorghum methylome by shallow sequencing of bisulfite-treated DNA. Through these efforts, we have established a strong association between various methylation patterns and high confidence gene models, and we have used these insights together with gene expression and sequence conservation to develop a classifier to predict functional structural annotations. We have also expanded the existing transcriptome atlas for sorghum by conducting RNAseq analysis on meristematic tissues, florets, and embryos, and these data sets have been used to improve on the existing community structural annotations. Finally, we have applied our newly developed classifier to the community Sbi1.4 (Paterson et al., 2009) and our novel GB2.0 annotations. Predicted gene models, classification data and results, and additional information are available on a github repository (https://github.com/warelab/sorghum-gene-classifier; accessed 29 July 2013).

Materials and Methods

Plant Material and Growth Conditions
All plant samples were collected from genotype BTx623. Tissues were collected, flash frozen in liquid nitrogen, and stored at −80°C until RNA extraction. Vegetative meristems (Stage 0) were excised using a dissecting microscope from ~29-d-old plants grown under glasshouse conditions (approximately 14 h light). Floral meristems, mostly of Stage 4 were excised from plants over a 15-d period (approximately 30–45 d after planting). Embryos were excised from immature seed during the milk stage of development (Gerik et al., 2003), whereas florets were stripped from immature panicles in the early boot stage (cream-colored florets and wispy panicles). The remainder of the sorghum tissues (hydroponically-grown roots and shoots; subjected to osmotic stress or exogenous ABA treatment) from which RNAseq and bisulfite sequencing were obtained, have been detailed previously (Dugas et al., 2011). For each tissue, subsamples from a given treatment (e.g., individual embryos) were pooled before RNA extraction, with each bulked sample of a given organ or tissue representing a biological replicate. Technical replicates were obtained by rerunning the same biological replicate cDNA preparation on a different lane of an Illumina flow-cell.

Transcript Profiling

RNA Isolation and mRNA Library Preparation
Following grinding of tissues in a mortar and pestle under liquid nitrogen, total RNA was extracted using miRNEasy kits (Qiagen, Valencia, CA). Each mRNA library was prepared according to the manufacturer’s instructions (mRNA-Seq sample preparation kit, Illumina Inc., San Diego, CA) using poly-A pull-downs and quality control was conducted as previously described (Dugas et al., 2011). The libraries were run on the Illumina GAIIx Sequencer using version 4 reagents for single-end non-strand-specific read sequences of 51 bp in length. These sequences have been deposited in GEO (Gene Expression Omnibus; GSE50464). To create a transcriptome atlas for sorghum, the reads from 23 RNAseq runs previously obtained from shoot and root tissue subjected to abiotic stress conditions (Dugas et al., 2011) were combined in silico with RNAseq reads from the 13 runs created from the organs or tissues detailed above (Table S1). Sequence files used for genome alignment were created by Illumina’s Real Time Analysis (RTA) software. A total of 696.5 million reads were generated that passed Illumina’s purity filtering and 634 million reads (91%) were aligned to the reference genome (see Table S1 for a breakdown by sample).

Transcript Assembly and Quantification

Figure 1 depicts a step-by-step flowchart showing the bioinformatic processes and/or tools and experimental inputs used in transcript assembly, gene model construction, and DNA methylation profiling. TopHat v1.0.13 (Kim and Salzberg, 2011; Langmead et al., 2009; Trapnell et al., 2009) was used to align RNAseq reads (FASTQ) to the BTx623 reference genome (release Sbi1.4 by Phytozone) using default parameters including allowing reads to align to multiple locations. The reads of each sample were aligned separately with TopHat, and resulting alignments were pooled by SAMtools (Li et al., 2009) before being processed by Cufflinks v2.0.0 (Trapnell et al., 2010) using default parameters to assemble potential transcriptional units. Cufflinks assembled 113,269 transcript fragments covering 58,392 non-overlapping loci, and these results were used in the gene annotation pipeline. Cufflinks was also used to estimate transcript abundance on a per sample basis. The gene level fragments per kilobase of transcript per million mapped reads (FPKM) values observed in any sample for Sbi1.4 high confidence genes were divided into quintiles: 1 = [0, 2.89249), 2 = [2.89249, 13.293), 3 = [13.293, 28.8215), 4 = [28.8215, 64.5828), and 5 = [64.5828, ∞). The expression levels of all gene models were mapped to these quintiles to facilitate our investigation of associations between expression and methylation.

Methylation Profiling

Whole Genome Bisulfite Libraries
To obtain an initial overview of DNA methylation across the sorghum genome, hydroponically-grown roots treated with 20 µM ABA or treated with 57 µM NaOH (carrier solution for ABA) were examined (see Dugas et al. (2011) for further growth condition detail). Flash-frozen root tissue was ground under liquid nitrogen, and high-quality
DNA was isolated utilizing the FastDNA SPIN Kit (MP Biomedicals, Santa Ana, CA USA). Library construction and bisulfite conversion were conducted as described previously (Regulski et al., 2013) and sequenced on Illumina GAIIX machines using PE50 protocol. Sequences have been deposited in GEO (GSE50464).

**Mapping DNA Bisulfite Sequence Data**

Illumina reads were trimmed to remove low quality bases at the 3’ end using the Burrow-Wheeler Aligner (BWA) script (Li and Durbin, 2009) followed by removal of adapter and linker sequences using the FASTX Toolkit. Reads were then aligned to the sorghum genome using the bisulfite sequencing specialization of RMAP that allows Ts in the reads to align to Cs in the reference (Smith et al., 2009). In addition, a maximum of 4 mismatches were permitted to allow for SNPs and sequencing errors. Sequences mapping equally well to multiple locations were removed, as were duplicate reads (i.e., PCR duplicates that start and end at the same position). Single-base methylation was quantified as the ratio of reads aligned over a given C in the reference genome that contain a C at that position versus the number of reads with either a C or a T. The conversion rate cannot be conclusively be determined as the conversion for chloroplast DNA is not available for root samples. However, false positive methylated cytosines are likely to be randomly distributed, and therefore not problematic for identifying hypomethylated regions.

**Characterization of Hypomethylated Regions in the Genome**

To divide the sorghum genome into contiguous segments of hypomethylated DNA, we applied a Hidden Markov Model (Durbin et al., 1998). This model includes a two-parameter transition matrix measuring the frequency of transitioning from hypomethylation (set at 0.0001) and the frequency of transitioning to hypomethylation (set at 0.00001), and this is applied for both CpG and CHG contexts (Regulski et al., 2013). The two probability parameters, pU and pM, correspond to the expected proportion of unconverted cytosine observed in a hypo- or hyper methylated region, respectively. The parameter pU is initialized as the mean proportion of C in BTx623 at C sites with an observed proportion less than 50%, and pM is initialized as the mean proportion of C in BTx623 at C sites with an observed proportion at least 50%. Finally,
the algorithm computes the posterior probabilities and determines the maximum likelihood parameters to update the transition and emission parameters (pU and pM) and iterates until the emission parameters converge to within a tolerance of 0.001.

**Gene Model Methylation Profiles**

Each gene model was assigned a methylation profile as follows. Upstream and downstream regions of 2 kb apiece flanking each gene were each divided into 5 bins of 400 nucleotides and the exonic portions were concatenated and divided into 10 equal sized bins. Intronic regions were excluded from this analysis to decrease the noise from transposons inserted into introns. Average methylation levels were calculated within each bin by dividing the number of reads with methylated Cs by the total number of reads (C + T) aligned to Cs in the reference. Empty bins, either with no Cs or no reads aligning to the Cs, were treated as null values when calculating average methylation across multiple genes. Average CpG, CHG, and CHH methylation plots were generated in R, a statistical programming language (R Development Core Team, 2008), to observe associations between methylation, expression, and gene length for high and low confidence genes (Fig. 2, discussed below). High confidence (i.e., ample transcriptional evidence as well as conservation across related species) and low confidence (often pseudogenes or models unsupported by transcription evidence) labels were assigned by the Sbi1.4 annotation (Paterson et al., 2009b); gene length was divided into long (at least 2 kb) and short (under 2 kb).

**Gene Structure Annotation and Comparison**

**Gramene/Ensembl Gene Annotations**

The Gramene/Ensembl GeneBuilder pipeline (Liang et al., 2009) was used to construct gene models from a variety of transcript evidence (Table 1, Fig. 3). In brief, the pipeline was run in three stages. Stage 1: we aligned to a soft-masked reference genome a compilation of evidence that included: transcript assemblies (Cufflinks output), proteins, mRNAs, and ESTs from sorghum and proteins and mRNAs from other monocot species. All available gene model evidence types, except RNAseq reads, were downloaded from Genbank as of May 6, 2010. The GeneBuilder pipeline uses Exonerate (Slater and Birney, 2005) to align various types of evidence to the reference genome. Stage 2: GeneBuilder then combined alignments from a specific evidence type to generate partial gene models. Stage 3: GeneBuilder combined these partial gene models to produce a working set of completed gene models. We treated the Sbi1.4 annotations as another partial models built during Stage 2 annotation. To assess the value of the RNAseq reads in gene model builds, two sets of gene models were constructed; one build that excluded the sorghum RNAseq assemblies (annotation version GB1.0) and a second gene model build that included them (annotation version GB2.0).

**Identification of Novel Transcriptionally Active Regions (nTARs)**

To identify nTARs, a methodology similar to that used for rice (Lu et al., 2010) was utilized. We excluded all genic regions identified in GB2.0 (± 2 kb) during nTAR screening as transcriptional activity in these regions may simply reflect incomplete gene models (e.g., incomplete UTRs or unannotated alternative splicing events). Stretches of contiguous expression in intergenic regions were identified by the following requirement: a read depth of 5 per bp for stretches of contiguous expression of 100 bp or greater or a read depth of 10 for a contiguous stretch of 50–100 expressed bp. Adjacent novel transcribed fragments separated by at most 10 bp were merged. To assemble transcribed fragments into multi-exon nTARs, the remaining intervals were compared to splice junctions found among the spliced Tophat alignments. Introns were created between fragments within 5 bp of a splice junction or between fragments separated by at most 2 kb with similar expression levels (read depth within 30%). To determine the potential for protein coding ability, nTARs were aligned to annotated proteins of maize, rice, and purple false brome [Brachypodium distachyon (L.)]. Compatible BLASTX (Altschul et al., 1990) hits between each protein and nTAR were merged to measure the extent of the protein coding region coverage by the nTAR. If this coverage was at least 50% of the protein’s length and the nTAR/protein alignment covered at least 50 bp of the nTAR, that nTAR was marked as possibly protein coding. nTARs were considered conserved, if at least 80% of the nTAR overlapped with whole genome alignments between sorghum and maize, rice, or Brachypodium.

**Analyzing and Comparing Sorghum Community and Gramene Gene Builds**

To discover differences between the Sbi1.4 community annotations and the GB2.0 gene models, we explored variations in gene structure of individual gene models and homology to other species. For homology-based comparisons, the canonical transcripts (the longest transcript with the longest coding sequence [CDS] from each gene model) were aligned to a blast database of annotated protein-coding genes from maize, rice, and Brachypodium. A significant change in homology was noted if the length of the aligned regions scaled by the percent identity changed by at least 10%. Using the Ensembl API (Potter et al., 2004), the lengths of the 5’ UTR, CDS, and 3’ UTR of each canonical transcript were calculated. Differences in their lengths were considered significant if the UTR length was changed by 10 bp or the CDS length changed by 3 bp. Overall, a gene model was considered improved if it exceeded one of these thresholds and didn’t decrease any of the other criteria by the aforementioned amounts.

**Phylogenetic Analyses of the Sbi1.4 Community Gene Annotations**

Comparative genomics resources provided by Gramene, release 34 (Youens-Clark et al., 2010) include phylogenetic
Figure 2. Cytosine methylation across annotated gene models and relationship between DNA methylation, transcription, and transcript length. Ends of gene models were aligned at the 5' end, and average methylation scores for each gene model (including 2 kb flanking regions) were split into 20 bins and plotted. The first five bins dissected the 2 kb upstream of the transcription start site (TSS), the next 10 bins dissect the exons, and the last five bins cover the 2 kb region downstream of the transcription termination site (TTS). The panels to the left show CpG (A), CHG (B), and CHH (C) methylation profiles of all Sbi1.4 genes on nuclear chromosomes divided into subsets based on confidence class (line type), transcript length (thickness), and expression level (color) in matched root tissues. Panels D, E, and F on the right show the corresponding methylation profiles (CpG, CHG, and CHH, respectively) of low confidence gene models that were classified as functional (solid lines) or non-functional (dashed lines) by our decision tree. Panels D, E, and F lack lines for non-functional long, highly expressed genes due to lack of data. For all six panels, the gene sets were bisected by expression level (threshold FPKM 2.89) and again by transcript length (threshold 2kb). Genes expressed above the threshold are turquoise whereas those expressed below the threshold are magenta. Genes longer than 2 kb are depicted by thick lines while thin lines depict shorter genes. The solid black lines represent average intergenic methylation levels for synthetic gene models for comparative purposes.
gene trees and whole genome alignments. As part of each Gramene release, the Compara pipeline (Vilella et al., 2009) is run to measure gene conservation across sequenced genomes of related grass species. We used Compara gene families to identify genes missing from the Sbi1.4 annotations, as well as split gene models that appear as adjacent partial genes in the same multiple alignment.

**Classification of Community Sbi1.4, GB2.0, and Novel Transcriptionally Active Region Models using Machine Learning**

To develop a classifier to distinguish between functional and non-functional gene models, we selected a subset of the Sbi1.4 high confidence and low confidence genes (as classified by Paterson et al. (2009b), as a training set. The “functional” gene models in our training set were the high confidence genes labeled “syntenic.” The “non-functional” gene models were those labeled as “pseudogene” or “transposon.” Successful machine learning requires a breadth of experimental data. We gathered gene expression, sequence conservation, and methylation data for each gene model. As measures of gene expression, we calculated (i) the maximum expression level (quintile) across all tissues or treatments and (ii) the breadth of expression (number of samples at quintile 3 or higher) for each gene model. As a measure of sequence conservation, we calculated the proportion of exonic bases overlapping with conserved blocks of DNA identified by Gramene’s whole genome alignment analysis (Youens-Clark et al., 2010). We also measured the proportion of exonic bases overlapping with CHG hypomethylated regions (HMRs), the intersection of CpG HMRs near the TSS (± 50 bp),

### Table 1. Gene model contributions by evidence type and per annotation method. NA, not applicable.

<table>
<thead>
<tr>
<th>Evidence type†</th>
<th>Sequences</th>
<th>Alignments</th>
<th>Partial models‡</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. bicolor</em> RNAseq</td>
<td>113,269</td>
<td>113,269</td>
<td>58,392</td>
</tr>
<tr>
<td><em>S. bicolor</em> protein</td>
<td>71,356</td>
<td>163,112</td>
<td>37,909</td>
</tr>
<tr>
<td><em>S. bicolor</em> mRNA</td>
<td>32,738</td>
<td>59,607</td>
<td>29,634</td>
</tr>
<tr>
<td><em>S. bicolor</em> cDNA</td>
<td>388</td>
<td>340</td>
<td>237</td>
</tr>
<tr>
<td><em>S. bicolor</em> EST</td>
<td>209,828</td>
<td>204,833</td>
<td>22,369</td>
</tr>
<tr>
<td>Monocot protein</td>
<td>459,854</td>
<td>565,280</td>
<td>24,498</td>
</tr>
<tr>
<td>Monocot mRNA</td>
<td>204,956</td>
<td>65,042</td>
<td>11,994</td>
</tr>
<tr>
<td>Monocot cDNA</td>
<td>79,005</td>
<td>31,835</td>
<td>10,851</td>
</tr>
<tr>
<td>Sbi1.4 models§</td>
<td>NA</td>
<td>NA</td>
<td>34,496</td>
</tr>
</tbody>
</table>

Annotation version | Gene models | Transcripts |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sbi 1.4</td>
<td>34,496</td>
<td>36,338</td>
</tr>
<tr>
<td>GB1.0</td>
<td>47,209</td>
<td>51,080</td>
</tr>
<tr>
<td>GB2.0</td>
<td>77,372</td>
<td>94,929</td>
</tr>
</tbody>
</table>

†Cufflinks assembled transcript fragments were aligned to the locations where they were defined, other evidence types were aligned with exonate.

‡Partial models were generated for each evidence type before being combined into the final set of gene models.

§Sbi1.4 gene models were treated as ab initio gene predictions by the GeneBuilder pipeline to combine compatible gene model fragments.

![Venn diagram](image-url)

**Figure 3.** Venn diagram depicting numbers of GB2.0 gene models supported by different evidence types in sorghum genome annotation. Diagram indicates the type of evidence and the number of gene models that were built from alignments of sequence data. The values given in parentheses represent the number of novel GB2.0 gene models supported by a particular combination of evidence not annotated in the published Sbi1.4 community annotation. The evidence type labels appear nearest the corresponding ellipse in a matching color.
and CpG HMRs at the 3’ end (−100bp). These six parameters were used to train a J48 decision tree classifier using the Weka suite of machine learning algorithms (Frank et al., 2004). The resulting decision tree (Fig. S1) was used to classify gene models from Sbi1.4, GB2.0, and the set of nTARs as potentially functional or non-functional. We also trained classifiers using support vector machines (SVM) with both linear and polynomial kernel functions and found that they have similar performance to the J48 decision tree with an F-measure of 0.933 on high confidence genes classified as functional and 0.699 on low confidence genes classified as non-functional. The decision tree is reported here because it offers an intuitive algorithm for sorghum geneticists to follow.

**Results**

**Genome-wide Patterns of Methylation and Gene Expression in Sorghum**

The structure of sorghum chromosomes is similar to that of rice and other members of Poaceae; specifically, gene-poor pericentromeric regions and gene-rich chromosome arms (Kim et al., 2005). Given this chromosomal structure, we examined the global RNAseq-based expression of Sbi1.4 annotated gene models and DNA methylation (CpG, CHG, and CHH) across the sorghum genome (Fig. 4).

As depicted in the circular plot of the 10 sorghum chromosomes, the euchromatic chromosome arms were hypomethylated (Fig. 4, innermost ring E), which correlates with the higher level of transcriptional activity associated with these genomic regions (Fig. 4, rainbow heat map, ring D). Similar patterns of hypomethylation were observed for cytosines in CpG, CHG, and CHH contexts with hypomethylation being a genomic signature of gene-rich chromosomal regions. In contrast, the pericentromeric regions (delimited by low gene expression and low hypomethylation signals) are rich in repetitive DNA (Fig. 4, ring C). This overall genomic pattern of transcriptional activity associated with the hypomethylated euchromatic arms is similar to what has been reported for rice and *Arabidopsis* (He et al., 2010; Lister et al., 2008).

**Gene-Space Methylation Patterns Associated with High- and Low-Confidence Sbi1.4 Gene Models**

To determine if characteristic patterns of gene-space cytosine methylation exist in sorghum, we visualized the different methylation patterns spanning Sbi1.4 gene models that were partitioned by length, expression level, and Sbi1.4 confidence class. Both the methylation patterns and expression levels of the genes were obtained from matched hydroponically grown root samples. In comparison to the intergenic average (represented as the black solid line in Fig. 2), high confidence gene models are, in general, hypomethylated (Fig. 2A-C). However, among the CpG, CHG, and CHH contexts, we observe distinct modulating patterns of cytosine methylation across the gene model; these distinctive methylation patterns were most apparent for highly expressed, high confidence gene models of at least 2 kb in length (Fig. 2A-C, solid, thick, turquoise line).

In examining CpG methylation, there was a distinctive pattern associated with all groups of high confidence gene models (solid lines) in which the transcription start site (TSS) and the transcription termination site (TTS) were hypomethylated (Fig. 2A). The central portion of gene bodies of some classes of gene models showed a higher percentage of CpG methylation that again declined when nearing the 3’ and 5’ transcript ends (Fig. 2A). This modulating pattern of CpG methylation status was more pronounced in long, highly expressed genes in both confidence classes (Fig. 2A, thick, turquoise lines). By contrast, low confidence gene models that displayed low levels of expression in either length class (Fig. 2A, magenta, dashed lines) did not show this increase in CpG gene body methylation pattern; rather, they displayed reasonably constant methylation of CpG sites (~50%) across the entire gene body and flanking sequences. Low confidence genes exhibiting higher expression did show hypomethylation at the TSS and TTS (Fig. 2A, turquoise, dashed lines), which may indicate that a subset of low confidence Sbi1.4 gene models may represent functional genes (see results below).

In examining gene methylation in the CHG context (Fig. 2B), all subsets of high confidence gene models were hypomethylated in the gene body whereas the 2 kb flanking regions showed greater methylation (~20%). The extent of hypomethylation of the transcribed region was most pronounced for highly expressed genes (Fig. 2B, turquoise, solid lines), although even weakly expressed high confidence gene models showed decreased gene body methylation compared to the 2 kb region flanking the gene body (Fig. 2B, magenta, solid lines). Similar to what was observed for CpG methylation, low confidence gene models (Fig. 2B, dashed lines) did not show a consistent pattern in CHG methylation status, although highly expressed gene models showed a reduction (~10–15% decreased CHG methylation) in gene body methylation when compared to regions upstream and downstream of the transcribed regions (Fig. 2B, dashed, turquoise lines). Similar to the CpG methylation, this CHG gene body methylation pattern could be associated with some low confidence gene models that may represent functional genes.

Gene body methylation in the CHH context was reduced in high confidence and highly expressed low confidence gene models, but the reduction was consistently low across the transcribed region in high confidence models whereas low confidence models showed apparent sporadic inflections in methylation status (Fig. 2C). These sporadic inflections could be an artifact of the extent of methylation sampling depth and require further experimentation to determine the authenticity of the spikes in methylation signal. Interestingly, as in maize (Gent et al., 2013; Regulski et al., 2013), an expression-associated peak of CHH hypermethylation was observed.
in a region approximately 400 bp upstream from the TSS. This hypermethylation peak was more pronounced for highly expressed high confidence (Fig. 2C, turquoise, solid lines) gene models reaching nearly 4% CHH methylation (compared to a genome average of ~2%). In general, this hypermethylated peak was not detected in low confidence gene models with the exception of highly expressed, long, low confidence gene models (again, indicating that these models may represent functional genes).

Given the lack of methylation data from other tissues, we can only partially explore correlation between expression and methylation. Hypomethylation seen at the 3' and 5' ends of high confidence genes is not correlated with expression in roots and shoots, whereas...
the increase in methylation in the gene body exons is obvious in genes expressed in both roots and shoots, but slight for root- or shoot- specific genes (data not shown). This suggests that broadly expressed genes are more likely to be methylated in the non-cis-regulatory regions. No significant differences in methylation patterns were observed based on gene expression in roots and shoots in other contexts (data not shown).

Developing a Functional Gene Classifier

In an attempt to develop new methylation-based criteria for use in a classifier to separate functional from non-functional gene models, the characteristic patterns of methylation in high confidence gene models reported in Fig. 2 were reassessed by comparing high and low confidence gene models with hypomethylated regions identified using a Hidden Markov Model (HMM). The most consistent methylation signatures (i.e., smallest standard deviations) across all annotated genes were the CHG-hypomethylated gene bodies and the CpG-hypomethylation coinciding with the 5’ and 3’ ends of transcripts (standard deviation data not shown). In contrast to the CpG and CHG hypomethylation patterns of high confidence gene models, the CHH hypermethylation signal is present in a relatively limited set of genes, and, thus, we did not include CHH methylation signals as a criterion in our gene model classifier. In general, most Sbi1.4 high confidence gene models, regardless of expression, exhibit two or more distinct methylation patterns. Greater than 73% of all high confidence gene models (20,195 of 27,640 models) displayed CHG hypomethylation of their exons and CpG hypomethylation at the TSS and TTS, while less than 5% (1275 of 27,640) of the high confidence models displayed only one gene space methylation pattern (Fig. 5A). Interestingly, although the vast majority did show two or more characteristic signatures of gene space methylation, 984 high confidence Sbi1.4 gene models showed no gene space methylation signatures. Therefore, it is important to include complementary data types such as gene expression and conservation for the classifier to utilize in these cases.

In examining the 6856 low confidence Sbi1.4 gene models, 47% (3216) showed no methylation signatures typified by high confidence gene models (Fig. 5B). However, a substantial number of Sbi1.4 low confidence gene models showed 3 or 4 signatures of gene model methylation. Nearly 23% (1557) of low confidence gene models showed CpG hypomethylation signatures at the TSS and TTS of transcripts along with CHG hypomethylation of gene bodies (Fig. 5B), which suggests that they may actually represent functional genes.

We developed and evaluated a gene model classifier to predict functionality based on gene body hypomethylation in the CHG context, CpG hypomethylation at the TSS and TTS, gene expression, and sequence conservation. A subset of the Sbi1.4 gene models were used to train a decision tree classifier (Fig. S1) that operates on six different features (see Methods). The classifier was then tested on the remaining genes, which have less convincing labels (i.e., mghNoSynteny, probableGene, and lowConfidence), as assigned in the Sbi1.4 annotation. The accuracy of the classifier, as defined by classifying high confidence genes as functional and low confidence genes as non-functional, on the training set was 97%, whereas its accuracy on the remaining Sbi1.4 genes was 74%.

To assess the contributions of methylation, expression, and sequence conservation data for predicting functional gene models, we trained and tested three additional classifiers each lacking one of the types of data. Among 27,444 high confidence gene models, 3113 (11.3%) had conflicting classifications between the partial classifiers. Each classifier predicted some genes as non-functional that the other two classifiers predict to be functional. These genes rely on the removed data type for correct classification. The classifiers trained without expression, methylation, and sequence conservation data each exclusively mispredicted 680 (21.8%), 1335 (42.3%), and 211 (6.8%) of these genes as non-functional, respectively. Among 6564 low confidence gene models, 706 (10.8%) had conflicting classifications with 9 (1.3%), 213 (30.2%), and 11 (1.6%) genes predicted as functional exclusively by the classifiers ignoring expression, methylation, and sequence conservation data, respectively. Each type of data contributes to the accurate classification of complementary subsets of gene models, with methylation data exclusively supporting the largest number of predictions. Therefore, combining all available data types will make for a more accurate and robust classifier.

Overall, the classifier identified 2520 (36.8%) low confidence gene models as functional and 1178 (4.3%) high confidence gene models as non-functional. The methylation profiles of the low confidence Sbi1.4 gene models classified based on our functionality prediction are shown in the second column of Fig. 2 (panels D, E, and F). The low confidence genes classified as non-functional approximate the CpG methylation levels of transposable elements (hypermethylated relative to background levels), whereas the various subsets of genes classified as functional (Fig. 2D) resemble the patterns observed for Sbi1.4 high confidence genes (Fig. 2A, solid lines). The functional Sbi1.4 low confidence genes (Fig. 2E, solid lines) all follow the characteristic pattern of CHG hypomethylation (Fig. 2B, solid lines) while those predicted to be non-functional are hypermethylated on average (Fig. 2B, magenta, dashed lines). Hypomethylation in the CHH context in the gene body is consistently low for genes reclassified as functional, whereas non-functional genes remain methylated at intergenic background levels (Fig. 2F).

Improvement of Sorghum Structural Annotations and the Impact of RNAseq Data

We used the Gramene-EnsEMBL GeneBuilder pipeline, which utilizes a compilation of transcript and protein information, both from sorghum and related species, to
build gene models and predict transcripts, with a focus on protein coding annotations (Liang et al., 2009). To assess the impact of RNAseq on gene models and predicted transcripts in sorghum, we utilized the Gramene-Ensembl GeneBuilder to generate two sets of gene builds; build GB1.0 excluded RNAseq data, and build GB2.0 incorporated the sorghum RNAseq atlas as an additional feature type. Considering the different forms of evidence separately, RNAseq transcript assemblies provided supporting evidence for the greatest number of gene model features.
affecting 55,077 GB2.0 models, while relatively few gene models (2154) depended on evidence from other monocot species (Fig. 3, orange-shaded ellipse). Overall, the GB2.0 annotation set contains 77,372 gene models. This represents an increase of more than 30,000 gene models due solely to inclusion of RNAseq data as a form of evidence (Table 1, GB1.0 vs. GB2.0). This increase includes various types of valid gene models, but also includes pseudogenes, transposable elements, anti-sense transcripts, and transcriptional noise, which artificially inflate the gene model count. The unfiltered gene model count of the RNAseq-based GB2.0 annotation represents a loose upper bound on the number of sorghum genes and shows the sensitivity of RNAseq and its potential impact on the fidelity of gene model predictions.

**Modification and Refinement of Sbi1.4 Gene Models**

It is recognized that the gene model prediction algorithms utilized during the Sbi1.4 community annotation and by the Gramene-Ensembl GeneBuilder used herein are unique and were applied to different sets of input data. To study the effects of these methods on the structure of gene models, we focused our comparison to the subset of GeneBuilder annotations that overlap with gene models in the Sbi1.4 published annotation. Of the 34,008 Sbi1.4 gene loci on assembled chromosomes, 89% (30,311) were concordant with GB2.0 gene models based on common genomic positions and 84% (28,619) were also concordant with GB1.0 models. The discrepancies between the total number of concordant gene models of the Sbi1.4 and GB2.0 annotations arise from three sources: (i) a set of 2818 Sbi1.4 split gene models that are merged in the GB2.0 annotation (see subsection below); (ii) 258 Sbi1.4 gene models that were split in GB2.0; and (iii) 621 of Sbi1.4 gene models that were absent from GeneBuilder annotations, which likely reflects slight differences in the gene building algorithms, especially with regard to the treatment and processing of repetitive elements.

The depth of the sorghum RNAseq data allowed us to reexamine Sbi1.4 gene annotation in a multitude of ways including evidence for extended exons, refinement of mapped 5’ and 3’ UTRs, the identification of alternatively spliced transcripts, and increased protein homology with monocot homologs. In examining the gene structure of GeneBuilder gene models and the Sbi1.4 annotation, there was no marked change in the median lengths of gene features or the ratio of single-to-multi-exon gene models (Table 2). However, there was a marked increase in GB2.0 gene models displaying alternative splicing events and extended UTRs (Table 2). The increase in alternative splicing events was partially due to differences in the algorithms used for gene model prediction (2.5-fold increase in GB1.0 vs. Sbi1.4 annotation), but the use of RNAseq transcript assemblies as evidence led to a much larger effect on predicted alternative splicing events (7.7-fold increase in the GB2.0 vs. Sbi1.4 annotation) (Table 2). We note that due to the complex nature of short read-based transcript assembly, the number of splicing events may be inflated and warrants additional curation.

Utilizing the RNAseq data, we quantified changes in the lengths of UTRs and coding regions as well as changes in homology when compared to proteins from maize, rice, or Brachypodium. Of the concordant gene models, 18,600 models (61%) showed an improvement in at least one category without a corresponding reduction in any other category (Fig. 6). Among these models, nearly 98% (18,216) have a UTR extended by at least 10 bp, 27% (5096) have a longer CDS, and 8% (1503) show at least a 10% improvement in homology to maize, rice, or Brachypodium (Fig. 6). In quantifying the impact of RNAseq on gene model characteristics, the largest change was observed in the UTRs, which increased on average by 141 bp, while the RNAseq data only resulted in an average increase of 4.8 additional amino acids per CDS and an average increase in homology of 0.96% (see Table S2 for details on specific directional changes). Finally, it should also be noted that a set of 3045 GB2.0 gene models (~10%) showed a 10 bp reduction in the length of either UTR, a shorter open reading frame (ORF), or 10% decreased homology coverage relative to the corresponding Sbi1.4 annotations (data not shown). Read depth coverage is not uniform across transcripts derived from the tissues or conditions sampled due to the technique used to generate RNAseq data, and, thus, may contribute to the observed reductions.

We examined in more detail the alternatively spliced gene models predicted by including RNAseq as evidence in GeneBuilder. Of those sorghum genes containing two or more exons, 33% produced two or more transcript isoforms supported by RNAseq (Table 2). Examining the different classes of alternative splicing,

---

**Table 2. Summary statistics of concordant gene models. UTR, untranslated region.**

<table>
<thead>
<tr>
<th>Gene model feature</th>
<th>Sbi1.4</th>
<th>GB1.0</th>
<th>GB2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-way concordant gene models</td>
<td>28,619</td>
<td>28,619</td>
<td>28,619</td>
</tr>
<tr>
<td>Single-exon models</td>
<td>7558</td>
<td>8512</td>
<td>7145</td>
</tr>
<tr>
<td>Multi-exon models</td>
<td>21,061</td>
<td>20,107</td>
<td>21,474</td>
</tr>
<tr>
<td>Alternatively spliced models†</td>
<td>916</td>
<td>2281</td>
<td>7112</td>
</tr>
<tr>
<td>Non-zero 5’ UTR</td>
<td>10,597</td>
<td>14,060</td>
<td>19,542</td>
</tr>
<tr>
<td>Non-zero 3’ UTR</td>
<td>10,520</td>
<td>13,810</td>
<td>19,721</td>
</tr>
<tr>
<td>Gene length‡</td>
<td>2040</td>
<td>1949</td>
<td>2371</td>
</tr>
<tr>
<td>Canonical transcript length*</td>
<td>1190</td>
<td>1160</td>
<td>1370</td>
</tr>
<tr>
<td>Canonical translation length‡</td>
<td>331</td>
<td>323</td>
<td>333</td>
</tr>
<tr>
<td>5’ UTR length‡</td>
<td>215</td>
<td>149</td>
<td>208</td>
</tr>
<tr>
<td>3’ UTR length‡</td>
<td>211</td>
<td>149</td>
<td>223</td>
</tr>
<tr>
<td>Cassette exon length‡</td>
<td>113</td>
<td>113</td>
<td>115</td>
</tr>
<tr>
<td>Intron length‡</td>
<td>145</td>
<td>145</td>
<td>146</td>
</tr>
<tr>
<td>Median intron count</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

†Alternatively spliced defined as a gene model displaying more than one predicted transcript isoform due to splicing events or alternative start or polyadenylation sites.

*Median lengths among non-zero length features are reported.
intron retention was the most common isoform (1606 models) while alternative 5’ donor and 3’ acceptor sites were observed in 671 and 1211 isoforms, respectively (Table 3). Exon skipping was observed in 576 RNAseq-supported gene models (Table 3). Gene structure analysis was conducted to determine the resulting coding region of the different isoforms, and the likelihood that the alternative splice events would result in transcripts that produce functional proteins was thus assessed. Of 7112 genes with multiple splice variants, 3949 (56%) contain a splice variant within 20% of the length of the canonical transcript (longest transcript with the longest translation), and of these, 3446 (87%) encode an altered coding region. It should be noted, however, that the authenticity of these isoforms remains to be determined (e.g., reverse transcription polymerase chain reaction [RT-PCR]) as does a determination if the isoforms produce stable transcripts or are targeted for nonsense-mediated mRNA degradation. In an effort to identify problematic RNAseq assemblies caused by non-stranded reads, we searched for gene models from Sbi1.4 and GB2.0 on opposite strands that overlap in exonic regions. There are 876 examples of GB2.0 gene models that overlap with multiple Sbi1.4 models from different strands. These cases tend to involve partial Sbi1.4 models with little support or nearby antisense pairs of genes whose 3’ ends overlap. The GeneBuilder produces models on both strands because it does not merge evidence specific gene models from opposite strands. There are also some examples of GB2.0 gene models predicted on the reverse strand opposite long exons. Thus, in the absence of stranded pair-ended RNAseq data or other evidence, further verification of the novel GB2.0 gene models and alternative isoforms by experienced curators is warranted before acceptance by the sorghum research community, as it is beyond the scope of this paper.

**Phylogenetic Analyses Reveal Missing and Split Gene Models**

To address missing gene models in the sorghum reference genome annotation, we utilized a phylogenetic approach across sequenced genomes of related grass species. Based on gene families generated using Compara, we identified 784 out of 22,522 gene families (3.5%) lacking a Sbi1.4 gene model but containing both a *Zea mays* gene and an *Oryza sativa* or *Brachypodium distachyon* (L.) gene. When sorghum genes identified in annotation GB2.0 were aligned to the Compara gene families lacking a Sbi1.4 gene model, 246 gene families were identified with an alignment covering at least 90% of one of the gene family members. The GB2.0 models that appear to be members of these 246

### Table 3. Different classes of alternative splice events supported by RNAseq reads in sorghum.

<table>
<thead>
<tr>
<th>Alternative splice event</th>
<th>No. of events</th>
<th>No. of genes¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intron retention</td>
<td>1986</td>
<td>1606</td>
</tr>
<tr>
<td>Altered 5’ donor</td>
<td>725</td>
<td>671</td>
</tr>
<tr>
<td>Altered 3’ receiver</td>
<td>1401</td>
<td>1211</td>
</tr>
<tr>
<td>Exon skipping</td>
<td>637</td>
<td>576</td>
</tr>
</tbody>
</table>

¹There are 2132 multi-transcript gene models with other types of differences, such as alternative 5’ and 3’ ends, mutually exclusive exons, multi-exon skipping, or other complex cases.
families include 192 updated Sbi1.4 models previously excluded from the Compara pipeline for insufficient protein length and 50 GB2.0 models that are novel relative to the Sbi1.4 annotation. Of the remaining 538 conserved gene families (representing 2.4% of annotated gene families) missing Sbi1.4 sorghum models, a large percentage (70%) showed limited homology (less than 90% coverage), suggesting a range of possible explanations including the sorghum homolog is a pseudogene in the reference genotype BTx623 or the reference genome sequence is incomplete. Nevertheless, over 99% of monocot gene families were annotated in the reference sorghum genome (see Table S3 for families not found).

We also utilized a phylogenetic approach to address the common genome annotation artifact of split gene models (i.e., annotation of two genes where phylogenetics suggest one should be), which result from fragmented genome assembly or gene models that are poor in supporting evidence. To search for this type of structural misannotation in the sorghum reference genome, we compared the Sbi1.4 reference annotation of sorghum to Compara gene tree models from maize, rice, and Brachypodium. Split sorghum gene models were identified by examining Compara gene trees for within-species paralogs in close genomic proximity that are disjoint in the multiple sequence alignment. Employing this phylogenetic approach, 427 putative structural misannotations were detected in the original Sbi1.4 reference annotation (Table S4, Fig. 7 shows example); of these, 51% (219/427) were corrected in the GB2.0 genome annotation. The reduced number of split gene models in the new GB2.0 genome annotation as compared to the community annotation results from additional evidence in sorghum (e.g., RNAseq) and related species since the release of the Sbi1.4 annotation in 2009. Of these “corrected” gene models, 45% (93/219) are likely to be complete annotated genes as they align over at least 90% of the length of another homolog from the gene tree. The remaining 208 predicted split gene models could have been missed for a variety of reasons including: false-positives arising from a gene fusion event in another species, insufficient evidence to assemble partial genes, misidentification of repeats, or missing genomic DNA in the BTx623 reference genome, missed either when sequencing or due to natural variation among sorghum genotypes.

**Novel Transcriptionally-Active Regions (nTARs)**

In addition to re-annotating existing protein coding gene models and identifying missing or split gene models, we searched for novel transcribed regions (nTARs) not predicted by the existing sorghum genome annotation or by our GeneBuilder pipeline. As GeneBuilder requires a gene model to contain an open reading frame of at least 50 amino acids, non-coding RNAs and genes encoding small peptides tend to be excluded. We identified 7359 nTARs, a large proportion of which (6704, 91%, data not shown) were shorter than 500 bp, and 81% display a single exon (Table 4). However, 1401 nTARs had multiple exons with an average length of 179 bp (Table 4 and data not shown). We detected 604 nTARs in regions of the genome conserved in maize, rice, or Brachypodium (Table 4). A few nTARs (77) were found to have a BLASTX alignment to a maize, rice, or Brachypodium protein covering at least 50 amino acids (Table 4). Recent interest in small proteins that play important regulatory roles in various biological processes warrants further characterization of nTARs containing short ORFs and the potential novel peptides they encode (Matsubayashi, 2011; Matsubayashi and Sakagami, 2006; Matsuizaki et al., 2010; Meng, 2012; Murphy et al., 2012; Wang and Fiers, 2010).

Utilizing a 1 Mbp sliding window to visualize the genomic distribution of nTARs (Fig. 4, red density plot; ring A) shows that nTARs are present throughout the euchromatic regions and absent from pericentromeric regions. The distribution of conserved nTARs (Fig. 4, red density plot; ring B) also follows the trend observed in protein-coding genes with a higher density in the euchromatic chromosomal arms.

**Classification of Novel GB2.0 Models and Novel Transcriptionally Active Regions**

Our classifier was applied to novel gene models identified in GB2.0 and the nTARs to assess potential functionality. Among the novel gene models (47,015 novel from GB2.0) and nTARs (7359), the decision tree classified 34,276
(63%) of them as potentially functional. The 30,555 novel GB2.0 gene models classified as functional are generally shorter than the Sbi1.4 high confidence models, especially in CDS length where the median translation is 25% as long (Table 5). There is a marked increase in the proportion of novel GB2.0 genes that are conserved between sorghum and other monocots when comparing the functional (43%) and non-functional predictions (22%). This measure of conservation is based on overlap between the gene models and whole genome alignment blocks. Aligning translations of novel gene models to SWISS-PROT/TrEMBL proteins from the Poaceae family identifies only 5725 genes with at least an 80% identity alignment covering at least 50% of the sorghum and SWISS-PROT/TrEMBL proteins. The large difference between the number of conserved novel gene models as measured by overlap with whole genome alignment versus alignment to a protein database is evidence of a protein-coding bias in the annotations of sequenced grass genomes. Therefore, we used another approach to study the conservation of the novel GB2.0 gene models that have open reading frames. The protein translations of these gene models and the Sbi1.4 gene models were aligned to the maize reference genome with exonerate using the protein2genome alignment model. The best alignments for each protein were checked for frame shifts and compared across four subsets; high and low confidence Sbi1.4 annotations, and novel GB2.0 models predicted to be functional or non-functional. We used the geometric mean of the percent similarity and the protein coverage to score alignments. As expected, the Sbi1.4 high confidence genes had the best alignments (median score 97%) and the fewest with frame shifts (6%). The low confidence genes had the poorest alignments (median score 87%) and the most frame shifts (17%) while the GB2.0

### Table 4. Comparison of gene features of distinct sets of gene models. aa, amino acids; CDS, coding sequence; nt, nucleotides; nTAR, novel transcriptionally active region.

<table>
<thead>
<tr>
<th>Gene model feature</th>
<th>Sbi1.4 high confidence</th>
<th>Sbi1.4 low confidence</th>
<th>nTARs</th>
<th>GB2.0 novel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>27,640</td>
<td>6856</td>
<td>7359</td>
<td>47,015</td>
</tr>
<tr>
<td>Single exon</td>
<td>5736</td>
<td>3084</td>
<td>5958</td>
<td>37,671</td>
</tr>
<tr>
<td>Multi-exon</td>
<td>21,904</td>
<td>3772</td>
<td>1401</td>
<td>9344</td>
</tr>
<tr>
<td>Alternative transcripts</td>
<td>1489</td>
<td>2</td>
<td>0</td>
<td>2737</td>
</tr>
<tr>
<td>Gene length†</td>
<td>2454 nt</td>
<td>701 nt</td>
<td>114 nt</td>
<td>438 nt</td>
</tr>
<tr>
<td>Transcript length†</td>
<td>1336 nt</td>
<td>489 nt</td>
<td>113 nt</td>
<td>413 nt</td>
</tr>
<tr>
<td>Protein coding</td>
<td>27,640</td>
<td>6856</td>
<td>77</td>
<td>47,015</td>
</tr>
<tr>
<td>Full length CDS‡</td>
<td>10,227</td>
<td>68</td>
<td>–</td>
<td>34,369</td>
</tr>
<tr>
<td>CDS length‡</td>
<td>359 aa</td>
<td>202 aa</td>
<td>–</td>
<td>63 aa</td>
</tr>
<tr>
<td>CHG HMR‡</td>
<td>25,843</td>
<td>2783</td>
<td>3226</td>
<td>30,474</td>
</tr>
<tr>
<td>CpG HMR 5′§</td>
<td>23,424</td>
<td>1849</td>
<td>1384</td>
<td>21,379</td>
</tr>
<tr>
<td>CpG HMR 3′§</td>
<td>21,068</td>
<td>1706</td>
<td>1375</td>
<td>20,794</td>
</tr>
<tr>
<td>Expressed¶</td>
<td>21,957</td>
<td>1379</td>
<td>5999</td>
<td>20,634</td>
</tr>
<tr>
<td>Conserved#</td>
<td>22,490</td>
<td>923</td>
<td>604</td>
<td>16,710</td>
</tr>
</tbody>
</table>

| Feature lengths given are median values.  
| Models with 5′ and 3′ untranslated regions (UTRs).  
| Overlap of at least 80% with target hypomethylated region (HMR). CHG, gene body exons; CpG, 50 nt before and after T5S; CpG3, last 100 nt of transcript.  
| Gene models with fragments per kilobase of transcript per million mapped reads (FPKM) > 2.89 in at least one tissue.  
| Gene models with at least 80% of their genomic span overlapping whole genome alignment blocks.  

### Table 5. Comparison of gene features of classified subsets of gene models. aa, amino acids; CDS, coding sequence; nt, nucleotides; nTAR, novel transcriptionally active region.

<table>
<thead>
<tr>
<th>Gene model feature</th>
<th>Functional low confidence Sbi1.4 models</th>
<th>Functional high confidence Sbi1.4 models</th>
<th>Functional nTARs</th>
<th>Functional novel GB2.0 models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>2520</td>
<td>1178</td>
<td>3721</td>
<td>30,555</td>
</tr>
<tr>
<td>Single exon</td>
<td>1253</td>
<td>242</td>
<td>2610</td>
<td>24,883</td>
</tr>
<tr>
<td>Multi-exon</td>
<td>1267</td>
<td>936</td>
<td>661</td>
<td>5672</td>
</tr>
<tr>
<td>Alternative transcripts</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1917</td>
</tr>
<tr>
<td>Gene length†</td>
<td>578 nt</td>
<td>1628 nt</td>
<td>144 nt</td>
<td>442 nt</td>
</tr>
<tr>
<td>Transcript length†</td>
<td>399 nt</td>
<td>975 nt</td>
<td>137 nt</td>
<td>418 nt</td>
</tr>
<tr>
<td>Protein coding*</td>
<td>2520</td>
<td>1178</td>
<td>60</td>
<td>30,555</td>
</tr>
<tr>
<td>Full length CDS‡</td>
<td>23</td>
<td>12</td>
<td>–</td>
<td>23,353</td>
</tr>
<tr>
<td>CDS length‡</td>
<td>244 aa</td>
<td>239 aa</td>
<td>–</td>
<td>61 aa</td>
</tr>
<tr>
<td>CHG HMR‡</td>
<td>2272</td>
<td>150</td>
<td>2961</td>
<td>22,806</td>
</tr>
<tr>
<td>CpG HMR 5′§</td>
<td>1767</td>
<td>51</td>
<td>1302</td>
<td>16,331</td>
</tr>
<tr>
<td>CpG HMR 3′§</td>
<td>1650</td>
<td>25</td>
<td>1300</td>
<td>15,902</td>
</tr>
<tr>
<td>Expressed¶</td>
<td>1154</td>
<td>96</td>
<td>2872</td>
<td>15,219</td>
</tr>
<tr>
<td>Conserved#</td>
<td>689</td>
<td>189</td>
<td>595</td>
<td>13,110</td>
</tr>
</tbody>
</table>

| Feature lengths given are median values.  
| Models with 5′ and 3′ untranslated regions (UTRs).  
| Overlap of at least 80% with target hypomethylated region (HMR). CHG, gene body exons; CpG, 50 nt before and after T5S; CpG3, last 100 nt of transcript.  
| Gene models with fragments per kilobase of transcript per million mapped reads (FPKM) > 2.89 in at least one tissue.  
| Gene models with at least 80% of their genomic span overlapping whole genome alignment blocks.  

OLSON ET AL.: EXPANDING AND VETTING SORGHUM GENE ANNOTATIONS 15 OF 20
novel gene models fall in the middle with better alignments and fewer frame shifts among those predicted to be functional (median score 96% and 14% frame shifts) than those predicted to be non-functional (median score 93% and 16% frame shifts). A supplementary figure (Fig. S2) shows the cumulative distributions of conservation with maize for these four sets of gene models. Among the novel GB2.0 gene models predicted to be functional 7654 (25%) are well conserved in maize with an alignment free of frame shifts scoring at least 90%.

The classifier, trained on syntenic protein coding gene models, transposons, and pseudogenes, was tested on known sorghum miRNAs to evaluate its usefulness for identifying functional non-coding genes. Our classifier predicted 92% of the miRNAs as functional, an encouraging sign that the model generalizes accurately. Further refinement of the classification of these gene models is beyond the scope of this paper. The sorghum community should consider the novel gene models as suggested functional loci, which may include, among other classes, partial protein coding genes, genes such as non-coding RNAs, or genes encoding short peptides. Repeating our efforts in a closely related species, such as maize, could validate these models and refine the classification criteria.

Discussion
As the de novo annotation of newly sequenced genomes and the refinement of existing genome annotations depend on the annotations of related species, continued refinement of the sorghum genome annotation will benefit not only the sorghum community, but also researchers working in related grass species. No additional refinement has occurred to the sorghum annotation since 2009, although new experimental data has been generated for sorghum and related monocot species. In particular, massive parallel cDNA sequencing (RNAseq) technology has been employed to characterize the sorghum transcriptome (Dugas et al., 2011). We utilized next-generation sequencing technology to generate a more comprehensive catalog of transcripts from sorghum to assist in the re-annotation of the sorghum genome including the identification of missing gene models and novel transcribed regions. We also developed a decision tree classifier to predict functional gene models based on sequence conservation, gene expression, and hypomethylation patterns observed from a 2x depth methylome sampling.

Heterochromatin Versus Euchromatin: Differences in Gene Expression, DNA Methylation, Density of Novel Transcriptionally Active Regions, and Repetitive Sequences
A previous study (Kim et al., 2005) determined that the sorghum chromosomes are characterized by relatively large pericentromeric regions of heterochromatin and arms that are largely euchromatic by nature. In the present investigation, the patterns of DNA methylation mirror the distribution of heterochromatin and repetitive DNA elements, which are concentrated in the pericentromeric regions of all 10 chromosomes. Consistent with DNA methylation’s role in repressing transcription of repetitive DNA, regions of the genome displaying high rates of transcriptional activity are hypomethylated and depleted of repetitive DNA (Fig. 4, D, E and C, respectively). Scattered across the chromosome arms are spikes of repetitive DNA and increased methylation status (Fig. 4, E and C, respectively), consistent with islands of repetitive elements scattered among gene-rich regions of the sorghum genome (Paterson et al., 2009a). As expected, the distribution of nTARs reflected the distribution of genome-wide gene expression with greater nTAR density on the euchromatic arms (Fig. 4, A and B).

Gene Body Methylation: Experimental Support for Functional Gene Models
Methylation of DNA is a heritable epigenetic signature that has been implicated in silencing the expression of genes, as evidenced by the high methylation status of transposable elements (Becker et al., 2011; Schmitz and Ecker, 2012). However, recent genome-wide scans in rice (Li et al., 2008; Su et al., 2011), Arabidopsis (Shen et al., 2012; Zilberman et al., 2007) and maize (Regulski et al., 2013) revealed characteristic signatures of gene body methylation that were correlated with gene expression levels.

In comparing high and low confidence gene models, as defined by the Sbi1.4 annotation, our data suggested characteristic signatures of gene body methylation that correlated with gene length and expression. High confidence gene models of all sizes and expression levels showed relatively high levels of cytosine methylation in the CpG context in regions 2 kb upstream or downstream of the gene body and hypomethylation of transcription start (5') and termination (3') sites. The lack of cytosine methylation at transcription initiation and termination and increased methylation in the gene body has been observed in other species including rice, Arabidopsis, and maize (Cokus et al., 2008; Lister et al., 2008; Regulski et al., 2013; Shen et al., 2012; Zilberman et al., 2007). It was proposed (Zilberman et al., 2007) that methylation impedes the initiation and elongation by Pol II and that methylation at the sites of transcription initiation and termination could result in stalling and collision of Pol II at these sites. Thus, highly expressed genes display a strong tendency to be hypomethylated at the ends of gene bodies. In addition to the hypomethylation signatures at ends of high confidence gene models, CpG methylation status increased at cytosines in the middle of the gene body with long, highly expressed genes showing the highest level of gene body CpG methylation. Methylation of CpG in the middle of the transcript is likely less problematic for Pol II processivity, although it is unclear why the extent of gene body methylation is positively correlated with gene expression level. In the CHG context, high confidence gene models are consistently hypomethylated.
relative to the flanking intergenic regions. In examining flanking DNA sequences, CHH methylation status of some highly expressed genes showed a hypermethylation peak 400bp upstream of transcription initiation. Methylation of upstream regions has been postulated to control transcriptional initiation at cryptic sites methylated via siRNA-directed machinery (Zilberman et al., 2007), but at present any relationship between the CHH hypermethylation peak and spurious transcription initiation in sorghum is purely speculative.

In the Sbi1.4 gene annotation, 5197 models were deemed low confidence due in part to lack of expression evidence (Paterson et al., 2009b). With the addition of 696.5 million sorghum RNAseq reads totaling 34.8 Gbp, we re-evaluated the classification of low confidence models. Nearly 900 low confidence gene models are expressed at second quintile or higher (FPKM > 2.89249) indicating their classification as low confidence should be reexamined in subsequent annotation updates. To this end, we employed a machine learning algorithm to build a decision tree classifier based on hypomethylation patterns, gene expression levels, and sequence conservation that identified a subset of 2520 Sbi1.4 low confidence gene models as likely functional.

Since our approach provided only a static overview of the methylation status of the sorghum genome, further experimental conditions (e.g., developmental stage, tissue type, environmental perturbation, genotypic effects) are warranted to fully characterize the methylome and its influence on gene regulation. As the sorghum genome community annotation continues the process of incremental updates, the methylation of gene models should provide additional evidence for bona fide gene models when used in conjunction with an expanding list of monocot RNAseq expression atlases.

Improving Sorghum Genome Annotation of Missing Models, Gene Model Structure, and Novel Transcriptionally Active Regions

By combining gene model evidence from RNAseq data, gene phylogeny, and existing forms of protein and transcript evidence, we generated a new annotation of the sorghum genome, annotation GB2.0. Compared to the community annotation version Sbi1.4, we see gene model improvements resulting from the incorporation of RNAseq data in the form of increased lengths of 5' and 3' UTRs in 18,584 and 18,861 existing gene models, respectively, and increased protein-coding sequence length or homology (i.e., protein identity with monocot homologs) for 19,036 gene models (Table S2). There was a subset of 18,332 genes (Table S2) that showed a reduction in gene model length or homology, suggesting that expert curation of structural changes to existing gene models is likely warranted. It should be noted that the changes in existing gene models were modest indicating that the published Sbi1.4 annotation is of high quality and only incremental improvements to existing gene models will be evident from the addition of RNAseq data. An exception to this may reside in the number of alternative splicing isoforms that were predicted in the published annotation and the number of isoforms suggested by RNAseq. Of the concordant Sbi1.4 gene models, 1360 were predicted with alternative spliced isoforms, whereas RNAseq-supported gene build predicted 7853 sorghum gene models with potential alternative splicing events (data not shown). The marked increase in the number of gene models showing alternative splicing isoforms may reflect the depth of coverage afforded by the present RNAseq data and the sensitivity of RNAseq to weakly expressed alternative splicing isoforms. A cautionary note must be issued related to the limitations of the present single-ended, non-strand specific RNAseq reads in relation to predicting alternative splicing events. The occurrence of natural antisense transcripts confounds the interpretation of alternative splicing events and potentially the accuracy of predicted gene boundaries, particularly at the 5' end. Thus, to build an accurate and precise map of gene boundaries and intron splice isoforms based on short reads, existing RNAseq data should be complemented with pair-ended strand-specific RNAseq (ssRNAseq) data and, ideally, confirmed by other forms of evidence (e.g., RT-PCR, Next-Gen Roche 454-based RNAseq analysis).

As more genomes become annotated, homology-based phylogenetic comparisons will reveal and aid in the correction of genome misannotations, which include missing and split gene models. By including RNAseq evidence during the GB2.0 gene build, the number of Sbi1.4 missing gene models was reduced from 784 to 548, whereas split gene models were reduced from 427 in the Sbi1.4 annotation to 219 in GB2.0 annotation. Hence, by employing a combined approach that exploits traditional gene model evidence from sorghum and related species and greater transcriptome coverage, this common gene structural error can be minimized in future genome annotations. As additional forms of evidence are included in future community gene builds, incremental increases in the number of annotated sorghum gene models will likely be evident rather than radical alterations in the total number of annotated protein gene models. Genotypic differences in gene content are expected as additional diverse genotypes of sorghum are sequenced and may help explain the vast phenotypic differences that exist amongst sorghums (Dahlberg, 2001).

In addition to providing evidence for structural modification of existing gene models, RNAseq also provided evidence for many novel gene models. Novel transcribed regions (nTARs) of a large number of genomes have been suggested, and the possibility that they encode regulatory molecules has spawned great interest in identifying and understanding their cellular function (Alexander et al., 2010; Brosnan and Voinnet, 2009). However, they have largely escaped detection in large-scale ab initio and evidence-based annotation of protein-coding gene models due to the criteria used. The present catalog of sorghum predicted novel GB2.0 gene models and nTARs constitutes a collection of previously uncharacterized proteins, long non-coding RNAs, and putative small peptides that may play important regulatory roles in
sorghum, such as mediating intercellular communication via mobile molecules (Matsubayashi, 2011; Matsubayashi and Sakagami, 2006; Matsuzaki et al., 2010; Meng, 2012; Murphy et al., 2012; Wang and Fiers, 2010). The expansion of genome annotation to include non-coding transcripts and those that potentially encode small peptides has already occurred for Arabidopsis, rice (Lamesch et al., 2012) and human (Harrow et al., 2012). According to the latest GENCODE statistics (version 17; http://www.gencodegenes.org/stats.html; Harrow et al., 2012), approximately 35% of the gene models in the human genome are protein-coding, 25% are pseudogenes, and 40% are non-coding RNA genes. Following experimental support by RT-PCR and mass spectroscopy, similar annotation of the sorghum genome with nTARs will be warranted.

Conclusions
Incorporating RNAseq data provided additional evidence for a more complete annotation of existing gene models in the Sbi1.4 community annotation of the sorghum genome, particularly those with little transcript support. RNAseq revealed many novel gene models and nTARs along with a genome-wide view of transcriptional activity in both heterochromatic and euchromatic regions of the genome. An initial view of the sorghum root methylese was obtained, which revealed characteristic signatures of gene model methylation and suggested a relationship between gene body methylation and gene expression at work in many genes. A classifier was developed that incorporated characteristic hypomethylation patterns complemented with measures of gene expression and conservation to predict gene models that are more likely to represent functional genes. The inclusion of this body of evidence into the community annotation of sorghum will benefit sorghum researchers and those working on related grass species.

Acknowledgments
The authors wish to thank Dr. Robert Martienssen (Cold Spring Harbor Laboratory) for bisulfite library support and discussions, Dr. Richard McCombie (Cold Spring Harbor Laboratory) for DNA sequencing support, Jude Kendall (Cold Spring Harbor Laboratory) for his technical assistance with HMMs, Joshua Stein (Cold Spring Harbor Laboratory) for assistance with comparative genomics analyses, and Julie McCollum (USDA-ARS, College Station) and Natalie Unruh (TAMU, College Station) for their expert technical assistance. This work was financially supported by USDA-ARS 1907-21000-030 (D.W.), NSF IOS-0703908 (D.W.), Texas AgriLife Research (P.E.K), and the USDA-ARS (R.R.K.).

References


Rewards in the plant genome. JULY 2014 ■ VOL. 7, NO. 2


Rewards in the plant genome. JULY 2014 ■ VOL. 7, NO. 2


Rewards in the plant genome. JULY 2014 ■ VOL. 7, NO. 2


Rewards in the plant genome. JULY 2014 ■ VOL. 7, NO. 2


Rewards in the plant genome. JULY 2014 ■ VOL. 7, NO. 2


Rewards in the plant genome. JULY 2014 ■ VOL. 7, NO. 2


Rewards in the plant genome. JULY 2014 ■ VOL. 7, NO. 2
