Sorghum (Sorghum bicolor) is a versatile C4 crop and a model for research in family Poaceae. High-quality genome sequence is available for the elite inbred line BTx623, but functional validation of genes remains challenging due to the limited genomic and germplasm resources available for comprehensive analysis of induced mutations. In this study, we generated 6400 pedigreed mutant pools from EMS-mutagenized BTx623 seeds through single-seed descent. Whole-genome sequencing of 256 phenotyped mutant lines revealed >1.8 million canonical EMS-induced mutations, affecting >95% of genes in the sorghum genome. The vast majority (97.5%) of the induced mutations were distinct from natural variations. To demonstrate the utility of the sequenced sorghum mutant resource, we performed reverse genetics to identify eight genes potentially affecting drought tolerance, three of which had allelic mutations and two of which exhibited exact cosegregation with the phenotype of interest. Our results establish that a large-scale resource of sequenced pedigreed mutants provides an efficient platform for functional validation of genes in sorghum, thereby accelerating sorghum breeding. Moreover, findings made in sorghum could be readily translated to other members of the Poaceae via integrated genomics approaches.
target genes that could be used to increase the quality and yield of grain and biomass. Once identified, such genes could then be modified via genome editing in crops with multiploid genomes (Bortesi and Fischer, 2015) by taking advantage of exciting progress on the RNA-guided CRISPR/CAS9 genomic editing system, which makes it possible to genetically improve organisms with complex genomes so long as the target genes are known (Jinek et al., 2012; Ran et al., 2013; Shalem et al., 2014). For example, brown midrib mutants increase the conversion efficiency of sorghum biomass to ethanol (Saballos et al., 2012; Sattler et al., 2012, 2014). As observed in sorghum, RNAi-mediated knockdown of the caffeic acid O-methyltransferase (bmr12) from sugarcane via RNAi increases fermentable sugar yield (Sattler et al., 2012; Jung et al., 2013). Therefore, discovery of gene mutations responsible for relevant traits in sorghum provides targets for genomic editing in other crops, thereby improving their utility for food and energy production.

EMS is a potent chemical mutagen that efficiently induces high-density mutations in genomes in many organisms (Greene et al., 2003). The development of TILLING (targeting induced local lesions in genomes) techniques, which can detect single-nucleotide mismatches in heteroduplexes of PCR amplicons, made EMS mutagenesis a workhorse technique for the study of recaclitrant metabolic and signaling processes in plants, animals, and microorganisms (Greene et al., 2003; Winkler et al., 2005; Gilchrist et al., 2006). In sorghum, a few mutated genes have been identified by TILLING based on individual amplicons (Xin et al., 2008; Blomstedt et al., 2012; Krothapalli et al., 2013). Recent advances in high-throughput next-generation sequencing have made it possible to catalog EMS-induced mutations on a large scale (Thompson et al., 2013; Henry et al., 2014). In rice (Oryza sativa) and wheat (Triticum aestivum), sequencing of captured exons has been used to identify large numbers of mutations from a limited number of mutant plants (Henry et al., 2014). However, this approach cannot capture mutations in regulatory regions that affect gene expression levels; moreover, it may not capture all exons with the same efficiency as whole-genome sequencing.

By linking large numbers of mutants with phenotypes of interest, crop plant mutation libraries have historically served as an important way to associate genes with their functions (Till et al., 2003; Settles et al., 2007; Krishnan et al., 2009). Previously, however, no comprehensive collection of sorghum mutant populations was available. To construct such a comprehensive mutant resource, we used whole-genome sequencing to discover mutations throughout the genome, as well as link mutations to phenotypes, in a pedigreed sorghum mutant library for which extensive phenotypic information is available (Xin et al., 2008, 2009, 2015). For this purpose, a mini-collection of 256 mutant lines randomly selected from the library was sequenced to an average coverage of 16x. Our results demonstrate that whole-genome sequencing of well-phenotyped mutant lines is an efficient way to identify causal mutations, as well as validate gene function by characterizing mutant phenotypes. Given sorghum’s relatively small diploid genome, low levels of gene duplication, and close phylogenetic relationship to maize and sugarcane, this resource should also facilitate the functional study of other grasses.

RESULTS

Pedigreed Mutant Library and Phenotypic Diversity

A pedigreed mutant library derived from a uniform genetic background is a powerful resource for linking genotypic variation to phenotypic diversity. We developed a mutant library in the sorghum inbred line BTx623, which was used to generate the sorghum reference genome. To generate the library, we introduced mutations using EMS and propagated individually mutagenized M1 seeds by pedigree to M3 seeds through single-seed descent (Figure 1). The mutants in the library exhibit a wide range of phenotypic variations, many of which segregate 1:3 (i.e., as Mendelian recessive traits) in M3 head rows. In addition to the variety of mutant phenotypes reported previously (Xin et al., 2008, 2009), we observed several traits in the mutant library that have the potential to be used for sorghum improvement (Table 1; Supplemental Data Set 1). For example, the multheaded (msd) mutants, which double the seed number and seed weight per panicle, could be used to improve grain yield (Burrow et al., 2014), and the brown midrib (bmr) mutants, mentioned above, could be used to improve biomass quality and ethanol conversion rate (Xin et al., 2009; Saballos et al., 2012; Sattler et al., 2012, 2014). Since the early 1930s (Duvick and Cassman, 1999), erect leaf architecture has contributed significantly to continuous yield gains in

![Figure 1](image-url)
maize hybrids; however, sorghum architecture has not been improved in this manner due to the lack of erect leaf germplasm (Assefa and Staggenborg, 2011). From our library, we isolated several erect leaf mutants (erf) that have the potential to improve sorghum canopy architecture to maximize biomass yield and stress resilience (Xin et al., 2009, 2015).

### Population Sequencing, Quality Control, and Mutation Detection

The wide phenotypic variations in this library prompted us to develop a high-throughput platform for rapidly identifying the causal mutations underlying these traits. A total of 256 M3 families were randomly selected for whole-genome sequencing. Because many M2 plants were sterile, genomic DNA was prepared from leaf samples pooled from 20 individual M3 plants in order to capture most of the mutations present in the original M2 plant. We obtained a total of 3.1 terabytes raw sequence data, which covered each genome at an average depth of 16.4x (range, 11 to 60x). Over 98% of short reads in each line could be mapped to the sorghum v2.1 reference genome, covering 86.6% of the genome and 95.6% of gene space (Supplemental Data Set 2). Genome coverage did not improve with increased sequencing depth from 11x to 60x (Supplemental Figure 1), probably due to the presence of repetitive DNA sequences and the difficulty of generating unique mappings from short reads (Sims et al., 2014).

To determine the proportion of mutations that were captured at the average sequencing depth (16.4x), we selected three lines with a sequencing depth of ~27x and compared the single-nucleotide polymorphism (SNP) discovery rate from 5x to 25x coverage. As shown in Supplemental Figure 2, 98% of homozygous SNPs and 85% of heterozygous SNPs that could be detected at 27x coverage were identified at 15x coverage. Thus, the vast majority of SNPs present in the mutant lines were captured at 16x genome coverage.

To validate the accuracy of the mutations identified using our bioinformatics pipeline, we randomly selected a total of 1024 SNPs (one heterozygous and three homozygous and SNPs from each mutant line) and subjected them to Sanger sequencing (Supplemental Data Set 4). Of the 1024 amplicons, 992 (96.8%) could be aligned to the targeted SNP regions, and 979 (95.6%) contained the predicted SNPs (Supplemental Table 1). The overall SNP accuracy was 98.7%/99.7% (767/769) for homozygous SNPs and 95.1% (212/223) for heterozygous SNPs. We divided the SNPs used for validation into five groups, according to the depth of sequencing coverage. SNP validation rates were similar (99%) for SNPs supported by 3 to 20 reads, whereas SNPs

<table>
<thead>
<tr>
<th>Mutants</th>
<th>No. of Lines</th>
<th>Potential application of Trait</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erect leaf (erf)</td>
<td>23</td>
<td>Biomass and grain yield</td>
</tr>
<tr>
<td>Brown midrib (bmr)</td>
<td>30</td>
<td>Biomass conversion efficiency</td>
</tr>
<tr>
<td>Multiple tiller (mtl)</td>
<td>120</td>
<td>Biomass yield</td>
</tr>
<tr>
<td>Early flowering (efl)</td>
<td>7</td>
<td>Biomass and grain yield</td>
</tr>
<tr>
<td>Late flowering</td>
<td>5</td>
<td>Biomass yield</td>
</tr>
<tr>
<td>Stiff stem (stl)</td>
<td>13</td>
<td>Biomass and grain yield</td>
</tr>
<tr>
<td>Early senescence (esn)</td>
<td>7</td>
<td>Early dehydration of biomass</td>
</tr>
<tr>
<td>Multiple seed (msd)</td>
<td>36</td>
<td>Grain yield</td>
</tr>
<tr>
<td>Large seed (lsd)</td>
<td>18</td>
<td>Grain quality</td>
</tr>
<tr>
<td>Bloomless (bmr)</td>
<td>107</td>
<td>Water-use efficiency</td>
</tr>
<tr>
<td>Heat sensitive (hs)</td>
<td>120</td>
<td>Tolerance to high temperature</td>
</tr>
</tbody>
</table>

The mutants were isolated over the last several years based on visual inspection of individual M3 head plots from a mutant library of 6400 lines. Some mutants segregated 1:3 in the M3 head plots, indicating they were recessive mutations in single nuclear genes. Identified mutants were crossed to a nuclear male-sterile mutant also isolated from the mutant library. The F1 and F2 plants were used to determine the genetic mode of the mutations.

We developed a rigorous bioinformatics pipeline to identify most of the EMS-induced mutations while minimizing the false positives. First, we removed all SNPs with more than two alleles. Furthermore, only canonical EMS-induced GC-to-AT transition SNPs were included in the analysis (Greene et al., 2003). To exclude background SNPs already present in our parental line before mutagenesis, we sequenced the BTx623 inbred line used to generate the mutant library to 12.9x coverage. We identified 13,243 high-confidence variations in our own parental line BTx623 relative to the sorghum reference sequence v2.1 (Supplemental Data Set 3). These SNPs occurred at high frequencies in the 256 lines we sequenced (Supplemental Figure 3); therefore, they are very likely to represent differences between our BTx623 line and the reference genome. Due to limitations on sequencing depth, this approach is unlikely to capture all background SNPs. Therefore, we amended the background SNPs with 65,634 variations that had allele frequencies >0.05 in the sequenced mutant population or that were detected in five or more mutant lines.

It is technically difficult to eliminate all contamination and cross-pollination during development of a large pedigree mutant library. Therefore, to ensure the quality of the mutation data, we first investigated whether any of the lines could have resulted from seed or pollen contamination. Two lines, ARS79 and ARS137, had extremely high apparent mutation rates (Supplemental Data Set 3 and Supplemental Figure 4A). Furthermore, the SNPs in these two lines overlapped extensively with natural variations (ARS79, 72%; ARS137, 77%; Supplemental Figure 4B). Therefore, it is likely that these two lines resulted from either seed contamination or pollination by unknown sorghum lines grown in the region; consequently, they were not included in subsequent analysis.

We also compared the SNP overlap between each pair of lines to identify possible siblings or cross-pollination among the mutant lines. Among the 64,516 pairs, two (ARS80/ARS85 and ARS20/ARS26) exhibited >50% similarity (Supplemental Figure 5A); ARS85 and ARS20, which had higher sequencing depths, were retained for downstream analysis. Following removal of these four lines, the identity of mutations between any two lines was <5% (Supplemental Figure 5B). After filtering the mutation database using the criteria described above, we identified 1,862,560 bona fide EMS-induced SNPs. On average, each sample contained 7660 SNPs (1798 homozygous and 5862 heterozygous; Supplemental Data Set 3). The average mutation density in the population overall was 11 SNPs/Mb, with a range of 8.02 to 22.5 SNPs/Mb (Figure 2A). A recent study of rice using exome capture technology reported a similar large range of variation in the mutation rate (Henry et al., 2014).
supported by 20 or more reads had a slightly lower validation rate, due to the heterozygous SNPs with high read depth. This slightly lower confirmation rate at high sequence depth might have been caused by paralogous alignment. The high SNP validation rate demonstrated that our variation call pipeline was robust, and the mutations discovered with our pipeline had high accuracy rate.

**Functional Annotation of the EMS-Induced Mutations**

Functional annotation of SNPs generated by EMS treatment was performed using the sorghum reference genome annotation v2.1. Based on the reference annotation, 22% of SNPs were located in gene space, covering ~95% (31,363) of genes (Table 2). We identified 4652 SNPs that changed an amino acid codon to a stop codon; 862 and 636 SNPs at splice acceptor and donor sites, respectively; and 86,070 missense mutations (Cingolani et al., 2012). Among the nonsynonymous SNPs, 29,120 had SIFT (sort intolerant from tolerant) scores < 0.05 and were therefore predicted to be deleterious mutations (Ng and Henikoff, 2003; Kumar et al., 2009). Overall, these 35,817 disruptive mutations covered 18,684 (57%) of annotated genes in the sorghum genome. On average, each line contained 147 large-effect mutations (stop-gain mutations, nonsynonymous coding mutations with SIFT score < 0.05 [Figure 2B] or mutations at splice acceptor or donor sites), with a range of 2 to 336 mutations per line (Figure 2C). Each line harbored an average of 145 genes with disruptive mutations (Supplemental Data Set 5). Furthermore, 28% (9157) of genes had disruptive mutations in more than one line, whereas 96% (8836/9157) of genes had unique mutations in two or more lines (Figure 2D; Supplemental Data Set 5). This finding provided a basis for identifying causal mutations defined by phenotypes with two or more independent alleles, allowing us to avoid the painstaking and costly approach of map-based cloning by classical forward genetics.

**Figure 2.** Characterization of EMS Mutations in 252 Sorghum Samples.

(A) Mutation density (SNP/Mb) in each sample.

(B) SIFT score distribution of the EMS mutations. Mutations with SIFT scores < 0.05 were considered to be deleterious to protein structure or gene function.

(C) Distribution of large-effect mutations in each sample. Large-effect mutations include stop-gain mutations, nonsynonymous coding mutations with a SIFT score < 0.05, and mutation at splice acceptor (two bases before exon start, except for the first exon) or donor sites (two bases after coding exon end, except for the last exon).

(D) Distribution of the number of large-effect mutations in each gene.
EMS-induced mutations (SNPs) were classified according to their location and potential effect on gene expression and protein structure. UTR, untranslated region.

Table 2. Annotation of Mutations Detected in the Sequenced Mutant Population

<table>
<thead>
<tr>
<th>Effect Type</th>
<th>SNPs</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>3' UTR</td>
<td>44,120</td>
<td>18,772</td>
</tr>
<tr>
<td>Start gained</td>
<td>5,237</td>
<td>4,326</td>
</tr>
<tr>
<td>5' UTR</td>
<td>29,452</td>
<td>14,055</td>
</tr>
<tr>
<td>Nonsynonymous coding</td>
<td>86,070</td>
<td>25,311</td>
</tr>
<tr>
<td>Start lost</td>
<td>136</td>
<td>136</td>
</tr>
<tr>
<td>Stop gained</td>
<td>4,652</td>
<td>4,043</td>
</tr>
<tr>
<td>Synonymous coding</td>
<td>138,416</td>
<td>22,035</td>
</tr>
<tr>
<td>Splice site acceptor</td>
<td>862</td>
<td>834</td>
</tr>
<tr>
<td>Splice site donor</td>
<td>636</td>
<td>621</td>
</tr>
<tr>
<td>Splice site region</td>
<td>6,894</td>
<td>5,260</td>
</tr>
<tr>
<td>Intron</td>
<td>195,916</td>
<td>21,443</td>
</tr>
<tr>
<td>Upstream 5 kb of genes</td>
<td>480,763</td>
<td>31,358</td>
</tr>
<tr>
<td>Downstream 5 kb of genes</td>
<td>464,091</td>
<td>31,141</td>
</tr>
<tr>
<td>Intergenic</td>
<td>1,457,491</td>
<td>–</td>
</tr>
</tbody>
</table>

Reverse Genetics Study Using Phenotype Data

To test the utility of our sequenced and phenotyped mutant lines, we focused on a frequently observed phenotype: bloomless (blm) mutants that lacked epicuticular wax deposition on their aerial surfaces. The blm mutants had well-defined shiny green sheaths that could be easily and unambiguously identified; moreover, this feature is potentially important in the context of drought tolerance and water-use efficiency (Burow et al., 2008).

Compared with other crop plants, sorghum accumulates a very high level of epicuticular wax on its aerial surfaces, which is believed to be critical to its high water-use efficiency, and tolerance to drought and heat (Burow et al., 2008). To date, however, no gene that contributes to epicuticular wax metabolism and deposition has been identified in sorghum. Very-long-chain fatty acid (wax; VLCF) metabolism has been well studied in Arabidopsis thaliana (Pighin et al., 2004; Samuels et al., 2008; Haslam et al., 2012) and is proposed to underlie epicuticular wax production in that species. The biosynthesis of VLCFs in Arabidopsis involves ketoacyl-CoA synthase (KCS), β-ketoacyl-CoA reductase, enoyl-CoA reductase, and several other genes related to biosynthesis of wax, such as CER1-5 (Pighin et al., 2004; Samuels et al., 2008; Haslam et al., 2012). To determine whether these genes are related to the wax pathway in sorghum, we searched for mutations in their sorghum putative orthologs. Six lines with the bloomless phenotype harbored mutations in eight sorghum orthologs of Arabidopsis VLCF metabolism (Table 3). The cer5 gene was represented by two allelic mutations in two different mutants. Fatty acid analysis of three blm lines (ARS20, ARS31, and ARS185) and the wild type confirmed that, as predicted, the three mutants lacked VLCFs. Moreover, the wax load was greatly reduced in blm mutants (Figure 3A).

We further validated the causal roles of these genes in the bloomless phenotype via two strategies: screening for independent allelic mutations and cosegregation test. A total of 30 blm mutants that were not included in the sequenced population were subjected to Sanger sequencing with overlap primers covering the eight sorghum orthologs of Arabidopsis VLCF genes. Six of the 30 mutants harbored distinct mutations in five of the eight genes. Two of the mutations were identical to those discovered by whole-genome sequencing of the 256 lines, whereas three were at different locations (Table 3). The detection of allelic mutations within genes with the same mutant phenotypes provides strong evidence for a causal link between gene and phenotype.

The ARS20 mutant line harbored heterozygous mutations in both the cer5 and cer6 genes. This unique line provided an ideal resource for cosegregation analysis between gene mutation and phenotype. Among 72 individual M4 plants obtained from a self-pollinated M3 plant (with heterozygous mutations in both genes), both cer5 and cer6 mutations followed an independent genotypic segregation ratio of 1 wild-type:2 heterozygote:1 mutant, whereas the bloom:bloomless phenotypic segregation was 3 bloom:1 bloomless. A goodness-of-fit test based on χ² analyses of both genotypic and phenotypic segregation ratios revealed that the observed values were statistically similar to the expected values (Figure 3B). In the population, only individuals with homozygous mutations in either or both of these genes exhibited a bloomless phenotype, consistent with the cosegregation pattern of the two genes (Figure 3C). These results indicated that the bloomless phenotype is a recessive trait, as expected for EMS-induced mutations. Each of the SNP mutations followed a monogenic pattern of segregation, and when we analyzed them together, we observed a pattern consistent with two genes segregating.

Table 3. Mutants of Sorghum Putative Orthologs in Arabidopsis Very Long Fatty Acid Synthesis Pathway That Exhibited the Bloomless Phenotype and the Corresponding Allelic Mutations

<table>
<thead>
<tr>
<th>Arabidopsis Gene</th>
<th>Sorghum Gene</th>
<th>Mutation Position</th>
<th>Amino Acid Change</th>
<th>Mutant</th>
<th>Independent Allele from Mutants Not in the Sequenced Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT1G85830 (CER6)</td>
<td>Sobic.001G453200</td>
<td>Chr1:655789406</td>
<td>E159K</td>
<td>ARS20</td>
<td>Chr1:655789403;A38T</td>
</tr>
<tr>
<td>AT2G28630 (KCS12)</td>
<td>Sobic.004G313000</td>
<td>Chr4:66583057</td>
<td>R189C</td>
<td>ARS73</td>
<td>Chr4:66583050;A38T</td>
</tr>
<tr>
<td>AT1G51500 (CER5)</td>
<td>Sobic.009G093300</td>
<td>Chr9:12576446</td>
<td>P581L</td>
<td>ARS73</td>
<td>Chr9:12578405;R188H</td>
</tr>
<tr>
<td>AT1G51500 (CER5)</td>
<td>Sobic.009G093300</td>
<td>Chr9:12579924</td>
<td>L244F</td>
<td>ARS20</td>
<td>Chr9:12578405;R188H</td>
</tr>
<tr>
<td>AT1G86530 (CER6)</td>
<td>Sobic.006G020600</td>
<td>Chr6:3500267</td>
<td>A133T</td>
<td>ARS20</td>
<td>Chr6:3500267;K284E</td>
</tr>
<tr>
<td>AT1G71160 (KCS7)</td>
<td>Sobic.002G268300</td>
<td>Chr2:65211661</td>
<td>E49S</td>
<td>ARS31</td>
<td>Chr2:65211661;K284E</td>
</tr>
<tr>
<td>AT1G19440 (KCS4)</td>
<td>Sobic.002G268500</td>
<td>Chr2:65229433</td>
<td>A49V</td>
<td>ARS31</td>
<td>Chr2:65229433;K284E</td>
</tr>
<tr>
<td>AT5G43760 (KCS20)</td>
<td>Sobic.005G168700</td>
<td>Chr5:55173639</td>
<td>R303Q</td>
<td>ARS185</td>
<td>Chr5:55173639;A49V</td>
</tr>
<tr>
<td>AT1G22205 (CER1)</td>
<td>Sobic.001G222700</td>
<td>Chr1:21258089</td>
<td>L100F</td>
<td>ARS185</td>
<td>Chr1:21258089;K284E</td>
</tr>
</tbody>
</table>
independently. As a negative control, two other heterozygous mutations in ARS20 were randomly selected for genotyping; their genotypes exhibited no cosegregation with the bloomless phenotype. Furthermore, independent alleles were identified for both cer5 and cer6 in the 30 blm mutants sequenced by Sanger sequencing (Table 3). Taken together, these cosegregation results provided strong genetic evidence that we had identified two independent genes (Sobic.009G083300 and Sobic.001G453200) and that these specific mutations were responsible for the bloomless phenotype in line ARS20. This example demonstrates that the sequenced sorghum mutant lines provide a useful resource for efficiently identifying mutations in genes in order to deduce their function through reverse genetics.

In Silico Analysis of QTLs and Mutations Using the Sorghum Mutant Resource

Identification of genes that underlie complex traits is challenging in many organisms, particularly crop plants. Especially problematic are the relatively small-effect quantitative trait loci (QTLs), which account for a large share of agriculturally important variations, as well as fitness differences in natural populations (Ellegren and Sheldon, 2008). We investigated whether our sequenced mutant library could help validate genes located in QTL regions. To this end, we focused on 13 seed size QTL intervals reported in four mapping studies (Paterson et al., 1995; Feltus et al., 2006; Srinivas et al., 2009; Zhang et al., 2015). Four of these 13 genomic regions had been identified in both QTL and GWAS studies. Of the eight genes flanking the GWAS peaks, four candidates (Sobic.004G136600, Sobic.006G268800, Sobic.007G166600, and Sobic.010G144900) were significantly associated with seed size, and two of these genes contained mutations in the sequenced mutants. ARS110 and ARS118, which harbor mutations in Sobic.006G268800, had higher proportions of large seeds and thousand-kernel mass (TKM) (Supplemental Table 2). ARS235 and ARS253, which harbor homozygous mutations in Sobic.004G136600, also exhibited increases in the proportion of large seeds and TKM; a putative ortholog of this gene was

Figure 3. Phenotype of Bloomless Mutants and Genetic Evidence.

(A) Characterization of bloomless mutants. Phenotype shows the presence or absence of epicuticular wax in stems of wild-type and mutant lines. The chromatogram from fatty acid analyses confirmed that C30 fatty acid, which is predominant in the wild type, was significantly reduced in bloomless mutants. (B) Scatterplots of endpoint genotyping results from an F2 population of ARS 20 for SbCER5 and SbCER6. (C) Genotype and phenotype data of F2 individual plants from ARS20. +/+ Wild type; cer5/+ and cer6/+, heterozygous mutants; cer5/- and cer6/-, homozygous mutants.
recently characterized in maize and rice in regard to its role in grain filling (Sosso et al., 2015). ARS37, which harbors a stop-gain mutation in Sobic.010G144900, had reduced seed size, similar to the small seed phenotype observed in several accessions from the Kafir working group that have spontaneous mutations in this gene (Zhang et al., 2015). This result showed that our resources could be very useful in identifying candidate genes and thus greatly aid in advancing QTL cloning research in crop plants.

Figure 4. Genome-Wide Distribution of EMS Mutations.
(A) Genomic coverage (sequencing depth) of the 252 sequenced lines.
(B) Mutation (SNP) density.
(C) Gene density.
(D) Karyotype.
Mutation Landscape in Sorghum

Our large-scale sequencing efforts revealed characteristics of EMS-induced mutations in sorghum. Henry et al. (2014) sequenced the exomes captured from 72 rice mutant lines and demonstrated that EMS preferentially induces mutations in sequences matching the consensus motif gggrRaR(g)CGrRgg (Henry et al., 2014). We analyzed the frequencies of the 20 nucleotides up- and downstream of 1.8 million G/C-to-A/T mutation sites and found that the proportion of C residues was elevated at the −2 and +1 positions relative to the mutation. This result is consistent with observations made in Escherichia coli, in which EMS-alkylated O6- guanidine residues near A/T pairs are preferentially removed via excision repair (Burns et al., 1986). Whole-genome sequencing of sorghum revealed no other obvious pattern, suggesting that EMS-induced mutation is largely random in this species (Supplemental Figure 4B). This conclusion is also supported by the nearly uniform distribution of mutations in all regions of the 10 sorghum chromosomes (Figure 4).

DNA methylation in sorghum did not appear to reduce the frequency of EMS mutation, as it does in rice (Henry et al., 2014). DNA methylation is distributed in a characteristic pattern across the chromosome, with more methylation in sequences closer to the centromere; by contrast, our EMS-induced mutations were mostly uniformly distributed (Figure 4). We evaluated the methylation level at EMS-induced mutation sites using previously published sorghum shoot DNA bisulfite sequencing data (Olson et al., 2013). The levels of methylation at both CpG and CHG (where H is A, C, or T) around EMS-induced mutation sites were higher than the average methylation level in the whole genome (Supplemental Figure 7A). In exon regions, methylation levels at CpG and CHG at EMS mutation sites were almost double the average level in the exome (Supplemental Figure 7B). This result indicated either that EMS prefers to alkylate G residues in regions with higher methylation rates or that a high methylation rate hinders the mechanisms that repair EMS-induced mutations in high-methylation regions (Burns et al., 1986).

DISCUSSION

Genetic variation, due to either natural or induced mutations, is the raw material for plant breeding. Recent studies showed that genetic diversity within cultivated crop varieties has been reduced by artificial selection during plant breeding (Fu and Somers, 2009). Consequently, mutagenesis represents an important source of artifi- cial selection during plant breeding (Fu and Somers, 2009). DNA methylation is distributed in a characteristic pattern across the chromosome, with more methylation in sequences closer to the centromere; by contrast, our EMS-induced mutations were mostly uniformly distributed (Figure 4). We evaluated the methylation level at EMS-induced mutation sites using previously published sorghum shoot DNA bisulfite sequencing data (Olson et al., 2013). The levels of methylation at both CpG and CHG (where H is A, C, or T) around EMS-induced mutation sites were higher than the average methylation level in the whole genome (Supplemental Figure 7A). In exon regions, methylation levels at CpG and CHG at EMS mutation sites were almost double the average level in the exome (Supplemental Figure 7B). This result indicated either that EMS prefers to alkylate G residues in regions with higher methylation rates or that a high methylation rate hinders the mechanisms that repair EMS-induced mutations in high-methylation regions (Burns et al., 1986).

A unique advantage of our mutant resource is that it provides the opportunity for swift integration of phenotypic data with the massive and detailed genomic sequence. As demonstrated here, functional genes and allelic series that control important agronomic traits (such as epicuticular wax production) can be identified and validated in mutant lines within a reasonably short period. Our cosegregation analyses of candidate genes (Sbcer5 and Sbcer6) for the bloomless trait in the M4–F2 population demonstrated that the mutant population could be used to resolve two independently segregating genes that affect a single trait. Thus, a translational genomics approach could be replicated with ease for candidate genes from model species such as Arabidopsis, especially for those involved in known biochemical pathways. The efficient validation of known genes in highly conserved pathways is an exciting and valuable application of the sorghum sequenced mutant resource.

In summary, we established a pedigreed mutant library in a sorghum inbred line, BTx623, that was used to generate the reference genome sequence. We sequenced a selection of 256 lines to an average of 16.4x coverage of the whole genome. The availability of this database will enable reverse genetics (in silico TILLING) to be conducted for a majority of the genes in sorghum. If additional alleles are needed for particular genes of interest, TILLING can be applied to an additional >6000 lines for which high-quality genomic DNA has been prepared. We demonstrated, through both simple genetics (b/m) and in silico analysis of candidate genes for QTLs, that this mutation database can be used to efficiently identify causal mutations. We believe that other traits could be analyzed with similar levels of success. Fast and efficient discovery of causal genes for agronomically useful phenotypes will accelerate the improvements of not just sorghum, but for other grasses in which there are limited avenues of forward and genetic research techniques due to genome complexity.

METHODS

EMS Treatment

EMS treatment was performed as described (Xin et al., 2008). Briefly, dry BTx623 seeds in batches of 100 g (~3300 seeds) were soaked with agitation for 16 h at 50 rpm on a rotary shaker in 200 mL of tap water containing 0.1 to 0.3% EMS (v/v). The treated seeds were thoroughly washed in ~ 400 mL of tap water for 5 h at ambient temperature; wash water was changed every 30 min. Mutagenized seeds were air-dried and prepared for planting. The air-dried seeds were planted at a density of 120,000 seeds per hectare. Before anthesis, each panicle was covered with a 400-weight rainproof paper pollination bag (Lawson Bags) to prevent cross-pollination. Each bag was injected with 5 mL Chlorpyrifos (Dow AgroSciences) at 0.5 mL/L to control maize (Zea mays) earworms that might hatch within the bag and destroy the seeds. Sorghum (Sorghum bicolor) panicles were harvested manually and threshed individually, and M2 seeds were planted as one row per head. To ensure high mutagenesis efficiency, only panicles that set 10% or fewer seeds were allowed to propagate. Three panicles from each M2 head row were bagged before anthesis. Only one fertile panicle was used to produce the M3 seeds. For DNA extraction, duplicate leaf samples were collected from the same fertile plant, and both the leaf samples and the panicle were barcoded. Seeds from the barcoded M2 plants were harvested as M3 families of seeds. For phenotypic evaluation, each M3 family of seeds was planted as one row in the field. Many of the mutant lines exhibited diminished seed production during the M3
genomic DNA sequence. The detected in BTx623, high-con genotypes in the population were discarded. After removing all SNPs de sequencing and alignment quality >20. Alleles present in more than two during library construction and sequencing. Variation calling was performed (http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Sbicolor) us com). All reads were Beckman Genomic Services, acquired by Genewiz (https://www.genewiz. Sequencing was performed on Illumina HiSeq 2000 sequencing system by

Variation Detection and Function Prediction

Sequencing was performed on Illumina HiSeq 2000 sequencing system by Beckman Genomic Services, acquired by Genewiz (https://www.genewiz. com). All reads were first aligned to the sorghum reference genome v2.1 (http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Sbicolor) using Bowtie2 (Langmead and Salzberg, 2012). The alignment files were converted to BAM format and sorted using Samtools (Li et al., 2009). Picard (http://broadinstitute.github.io/picard/) was used to filter duplicated reads during library construction and sequencing. Variation calling was performed with Samtools and Bcftools (Li et al., 2009) using unique reads with both sequencing and alignment quality >20. Alleles present in more than two genotypes in the population were discarded. After removing all SNPs detected in BTx623, high-confidence SNPs were selected using the following criteria: (1) the SNP is supported by at least three reads or two complementary reads; (2) the mutation is detected in no more than five lines with an allele frequency ≤5%; and (3) the sequence change is GC→AT.

To validate a SNP, primers were designed based on the 2 kb of flanking genomic DNA sequence. The “primer3” module from EMBOSS (Rice et al., 2000) was used to find the primers with the following parameters: “-target 600.700 -psizeopt 600 -prang [500-700].” The copy number of primers was checked by searching the whole-genome sequence using BLAST (Altschul et al., 1990). Only single-copy primers were used for Sanger sequencing. The Sanger signals were converted to sequences using the SangerR (Poly Peak Parser) package (Hill et al., 2014). Heterozygous SNPs were identified based on ratio of peaks between two alternative nucleotides (Supplemental Figure 8). If the second-strongest peak signal was greater than one-third of the strongest peak signal, the SNP was scored as heterozygous (Hill et al., 2014). Amplicon sequences were aligned to the sorghum reference genome and then variations were called using the MUMMER package (Kurtz et al., 2004).

Functional annotation of SNPs was performed using Snpeff (Cingolani et al., 2012) based on the sorghum genome annotation v2.1. SIFT 4G was used to predict deleterious mutations (Vaser et al., 2016).

Characterization of Sorghum EMS-Induced SNPs

A sliding-window method (100 kb window; 10-kb step size) was used to determine the whole-genome distribution of EMS SNPs. For methylation analysis, published sorghum shoot bisulfite sequencing data were used (Olson et al., 2013). All reads were aligned to the sorghum reference genome v2.1 using Rmap (Smith et al., 2008), allowing a maximum of four mismatches. Only sites with more than five unique mapped reads were used to estimate the methylation level.

To determine the novelty of EMS-induced SNPs from natural variations annotated based on published literature (Mace et al., 2013; Morris et al., 2013), it is necessary to ensure that the EMS-induced SNPs and natural variations are at the same locations. Because all natural variations have been analyzed using the v1.4 sorghum genome annotation, the genes from the v1.4 annotation were first aligned to the v2.1 assembly using BLAST (Altschul et al., 1990). Only genes with a 100% match to the same chromosome from the v2.1 genome assembly were used for the SNP comparison. EMS-induced SNPs were compared with two natural variation sets available from the Gramene database (Monaco et al., 2014): The first population consists of 265,000 SNPs from 971 worldwide accessions, assayed by genotyping-by-sequencing (Morris et al., 2013); the second population consists of six million SNPs, assayed by whole-genome sequencing of 45 S. bicolor and two S. propinquum lines (Mace et al., 2013). The coordination of SNPs within the mapped genes was converted from the v1.4 assembly to v2.1 using a script written in house. Overlaps between VCF files were detected using BEDTools (Quinlan and Hall, 2010).

Phenotyping of Epicuticular Wax Mutants and Gas Chromatography

The 256 mutants were planted in the field for visual inspection in two consecutive years (2013 and 2014) and in a greenhouse (2015) at the experimental facilities of the Cropping Systems Research Laboratory, USDA-ARS. Mutant lines that were devoid of epicuticular wax (EW) were isolated, photographed, and subjected to EW analysis. EW from leaf and stem was extracted by drying a cut sample of ~6 cm² in 6 mL of hexane for 60 s. Samples were dried under nitrogen, subjected to esterification using the MIDI protocol, dried again, and reconstituted in hexane containing internal standard (http://www.midi-inc.com). Wax analysis was performed on a gas chromatograph coupled with a flame ionization detector.

Cosegregation Test of the Bloomless Mutant Population

An M4 population of ARS20 obtained from M3 plants (with heterozygous mutation in both cer5 and cer6) was used as an F2 population to analyze the cosegregation between mutation and phenotype. Eighty M4 seeds of ARS20 were planted in the polyhouse under optimal conditions (12-h/12-h day-night light regime with temperatures of 24/20°C, day-night conditions; 50% RH) and provided with automatic irrigation. After 1 week, 72 seeds out of 80 germinated and were used for the study. Individual M4 (F2) plants were tagged and the bloom/bloomless phenotype was scored after 30 d for each of the lines in the population and continuously monitored throughout the plant cycle. Bloom plants produced profuse white EW in the stem, while bloomless plants are devoid of EW with shiny green stem color specifically at the four-leaf stage (30 d after planting).

Genomic DNA was extracted from each of the individual F2s using the CTAB method referenced above. Allele-specific genotyping was performed by endpoint analysis using Kompetitive Allele-Specific PCR (KASP) chemistry (LG C Genomics). In-house primers were designed and synthesized for the mutations in Sobic.009G083300 (SbCER5) and Sobic.001G453200 (SbCER6) and two negative control loci (Sobic.004G086300 and Sobic.006G127800) (Supplemental Table 3). The PCR conditions for KASP analysis of sorghum were adapted from Chopra et al. (2015). χ² analyses were performed to assess the goodness of fit between the observed and expected values of the segregation ratio.

In Silico Analysis of QTLs and Mutations in Seed Size

Genes associated via bi-parental mapping or GWAS with QTLs for seed characteristics were identified in rice (Oryza sativa), maize, and sorghum. These genes were further assessed for the presence of mutations in whole-genome shotgun sequences of the sequenced lines. Seed samples from 63 selected mutant lines and BTx623 (wild type), grown in two successive years, were obtained based on whether a specific line harbored a mutation or SNPs in previously reported seed characteristic-related QTLs for sorghum, rice, or maize. To determine variations in seed diameter, the total harvested seed lot from each selected line was passed through a series of three sieves with 3.35 mm (#5), 2.8 mm (#6), and 2.45 mm (#7) screens. Percentages of total kernels retained in each screen were recorded, and average TKM was calculated for each of the selected mutant lines.
SNPs associated with seed size, different seed diameter classes and total kernel weight were statistically compared with those of the wild type.

Accession Numbers

All sequencing data have been deposited in the Sequence Read Archive (SRA) database under accession number SRP063947. The mutation data can be found at the Gramene database (http://ensembl.gramene.org/Sorghum_bicolor/Info/Index).

Supplemental Data

Supplemental Figure 1. Average sequencing depth and genome coverage in 256 samples.

Supplemental Figure 2. SNP discovery rate at various sequencing depths.

Supplemental Figure 3. Allele frequency of BTx623 SNPs and the NON-BTx623 SNPs in the 256 samples.

Supplemental Figure 4. Detection of two contaminating lines.

Supplemental Figure 5. Detection of sibling lines.

Supplemental Figure 6. Distribution of flanking nucleotides.

Supplemental Figure 7. Methylation profiling of EMS mutation.

Supplemental Figure 8. Validation of heterozygous mutations.

Supplemental Table 1. SNP validation rates at various sequencing depths.

Supplemental Table 2. Mutations associated with seed-size QTLs.

Supplemental Table 3. Primer sequences used for genotyping F2 population of ARS20.

Supplemental Data Set 1. Mutant phenotypes observed in each of the 256 sequenced lines.

Supplemental Data Set 2. Sequencing summary of 256 mutants.

Supplemental Data Set 3. Number of SNPs and mutation rate in each line.

Supplemental Data Set 4. Primers designed for SNP validation by Sanger sequencing.

Supplemental Data Set 5. Disruptive SNPs in each line.

ACKNOWLEDGMENTS

We thank Lan Liu-Gitz and Halee Hughes (USDA-ARS) for technical support and Veronica Acosta-Martinez (USDA-ARS) for technical assistance with GC analysis of waxes. The project was funded by ARS in-house project 6208-21000-020-00D and the United Sorghum Checkoff Program. We also acknowledge the funding from National Science Foundation IOS 1127112 (Y.J. and B.W.) and USDA-ARS 1907-21000-030-00D (D.W.) for data analysis. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA.

AUTHOR CONTRIBUTIONS

Z.X., J.B., and D.W. conceived the idea. Y.J. and Z.X. drafted the manuscript. Z.X. selected the mutant lines and prepared the genomic DNA. Y.J. and B.W. analyzed the data. R.C. and G.B. conducted reverse genetics and integrated genomics analyses of the bloomless and seed size mutants and genes. J.C., G.B., C.H., and Y.E. assisted with phenotyping at various stages. All authors edited and agreed on the final manuscript.

Received May 6, 2016; revised June 23, 2016; accepted June 23, 2016; published June 27, 2016.

REFERENCES


A Sorghum Mutant Resource as an Efficient Platform for Gene Discovery in Grasses
Yinping Jiao, John Burke, Ratan Chopra, Gloria Burow, Junping Chen, Bo Wang, Chad Hayes, Yves Emendack, Doreen Ware and Zhanguo Xin

Plant Cell 2016;28;1551-1562; originally published online June 27, 2016;
DOI 10.1105/tpc.16.00373

This information is current as of July 31, 2017

Supplemental Data
/content/suppl/2016/06/27/tpc.16.00373.DC1.html
/content/suppl/2016/09/19/tpc.16.00373.DC2.html

References
This article cites 67 articles, 20 of which can be accessed free at:
/content/28/7/1551.full.html#ref-list-1

Permissions

eTOCs
Sign up for eTOCs at:
http://www.plantcell.org/cgi/alerts/ctmain

CiteTrack Alerts
Sign up for CiteTrack Alerts at:
http://www.plantcell.org/cgi/alerts/ctmain

Subscription Information
Subscription Information for The Plant Cell and Plant Physiology is available at:
http://www.aspb.org/publications/subscriptions.cfm