Evaluation of Genetic Variation among Sorghum Varieties from Southwest China via Genome Resequencing

Songxian Yan,* Li Wang, Liang Zhao, Heyu Wang, and Diqiang Wang

ABSTRACT Little is known regarding genomic variation among glutinous sorghum \textit{(Sorghum bicolor (L.) Moench)} varieties grown in southwest China, which are primarily used to brew the popular Jiang-flavor liquor. This study evaluated genomic variation among six representative sorghum accessions via whole-genome resequencing. The evaluation revealed 2,365,363 single-nucleotide polymorphisms (SNPs), 394,365 insertions and deletions, and 47,567 copy number variations among the six genomes. Chromosomes 5 and 10 showed relatively high SNP densities, whereas whole-genome diversity in this population was low. In addition, some chromosomal loci exhibited obvious selection during the breeding process. Sorghum accessions from southwest China formed an elite germplasm population compared with the findings of other geographic populations, and the elite variety ‘Hongyingzi’ contained 79 unique genes primarily involved in basic metabolism. The six sorghum lines contained a large number of high-confidence genes, with Hongyingzi in particular possessing 104 unique genes. These findings advance our understanding of domestication of the sorghum genome, and Chinese sorghum accessions will be valuable resources for further research and breeding improvements.

Abbreviations: CNV, copy number variation; KEGG, Kyoto Encyclopedia of Genes and Genomes; SNP, single-nucleotide polymorphism; WGS, whole-genome shotgun.

CORE IDEAS

- Sorghum accessions from southwest China formed an elite germplasm population.
- The variety ‘Hongyingzi’ contained 79 unique genes and had a large planting area.
- The identified molecular markers will provide useful tools for crop improvement.

Sorghum is an important cereal crop worldwide because of its good adaptability and diverse uses. It has been widely accepted that sorghum cultivars originated from wild accessions in Africa (Mace et al., 2013) and that the availability of current varieties was achieved via domestication, which initially began 10,000 yr ago in northeast Africa. Domesticated varieties subsequently diffused to other regions in Africa, southwest Asia, the Middle East, and the rest of the world (Grenier et al., 2004; Morris et al., 2013). In addition to Sudan, an important center of sorghum domestication, China, Ethiopia, and India are considered to be centers of diversity for sorghum (Ayana and Bekele, 2000; Grenier et al., 2004; Li and Li, 1998). Sorghum has been adapted to different environments, which has led to the appearance of diverse morphological and physiological traits (Deu
et al., 2006; Upadhyaya et al., 2009) and, subsequently, crop improvement. As one species of sorghum, S. bicolor subsp. bicolor includes five basic races depending on the spikelet or panicle morphology, namely bicolor, guinea, durra, caudatum, and kafir, which exhibit no reductive isolation between each other (Dillon et al., 2007; Harlan and deWet, 1972). A large number of local accessions and cultivars are available in different regions worldwide. For example, more than 120,000 Chinese sorghum landraces have been collected in the National Gene Bank of China (Li and Li, 1998).

Sorghum bicolor has a long cultivation history in China; it is currently the fifth most cultivated cereal crop in China. Sorghum bicolor is mostly grown in lean and arid regions and can be used as food, feed, liquor, and fuel. In the process of adaptation to diverse environments, a rich resource of sorghum germplasm has been formed. Sorghum (known as kaoliang in China), is the raw material for some of the most famous liquor factories; it has been playing a key role in brewing liquors since ancient dynastic periods (Wang et al., 2015). For example, Moutai liquor (China Kweichow Moutai Distillery Co., Ltd., Renhuai City, Guizhou, China) is one of the most famous distilled beverages, similar to Scottish whiskey and French brandy (Zhu et al., 2007); it plays an important role in Chinese culture and the daily lives of Chinese people. Glutinous sorghum harvested from the local environment is a particularly important raw material for elite Moutai liquor. The starch in grains provides the glucose and maltose source for microorganisms after being degraded by some fungal species (Wu et al., 2009). Because of the high proportion of amylpectin (>88%) in sorghum grains, a high gelatinization temperature is required during fermentation; this provides a sustained sugar source throughout all seven cycles of raw liquor fermentation. Another characteristic of the glutinous sorghum is the high tannin content (1.5–2.0%). Tannins are located mainly in the testa and pericarp of the sorghum caryopsis (Hahn and Rooney, 1986; Taylor et al., 2013); they can restrict the excessive activity of amylolytic enzymes in each fermentation cycle of the popular Jiang-flavor liquor (also known as “sauce aroma”) and provide an important source of aromatic compounds (Wang et al., 2016). The Jiang-flavor liquor industries in southwest China have developed rapidly over recent years, which has led to the formation of the elite germplasm of glutinous sorghum from headstream.

As a crop with high photosynthetic efficiency (C₃), S. bicolor has a small genome of approximately 730 Mb (Paterson et al., 2009) and a simple genetic mechanism (2n = 20), making it convenient for studying genome function. Previous studies have indicated that the sorghum germplasm in centers of domestication, especially the African center, has rich genetic diversity (de Alencar Figueiredo et al., 2008; Grenier et al., 2004; Hariprasanna and Patil, 2015; Tesso et al., 2008). For example, Sudanese sorghum landraces have high diversity in terms of their growth period and kernel color, and they are distributed in different geographical regions (Grenier et al., 2004).

Sorghum accessions from Sahelian Africa also exhibit many genetic variations, whereas those from South Asia have similar diversity (Folkertsma et al., 2005). Conversely, Chinese sorghum landraces experienced severe selection pressure during the domestication process, resulting in low genetic diversity (Li et al., 2010). Thus genetic markers could be easily applied to cluster different evolutionary groups (Bouchet et al., 2012), create consensus genetic maps (Mace et al., 2009; Menz et al., 2002), and analyze genome-wide traits (Bekele et al., 2013; Kong et al., 2014; Thurber et al., 2013; Zhang et al., 2015). Moreover, whole-transcriptome sorghum microarrays have been used to identify tissue- and genotype-specific expression patterns of S. bicolor exons and untranslated regions (Shakoor et al., 2014).

Diverse modalities have been developed for plant genome sequencing, such as bacterial artificial chromosome libraries, whole-genome shotgun (WGS) strategies, and next-generation sequencing (Goff et al., 2002; The International Barley Genome Sequencing Consortium, 2012; Tuskan et al., 2006; Varshney et al., 2009; Xu et al., 2013; Yu et al., 2002). The S. bicolor genome was sequenced via a WGS approach in 2009 (Paterson et al., 2009), providing a model for studying the genomic domestication and genetic variations of this multipurpose crop. Researchers have identified numerous SNPs, insertions and deletions (indels), and present or absent polymorphisms in the sorghum genome via whole-genome resequencing (Bekele et al., 2013; Mace et al., 2013; Shen, Liu et al., 2015), and these high-throughput molecular markers are useful for genome-wide association studies. Meanwhile, some nucleotide-binding site-encoding genes in the sorghum genome are enriched through different processes of domestication and improvement (Mace et al., 2014; Zheng et al., 2011).

Although several reports have described genetic variations among different geographical sorghum accessions or breeding populations, little is known about genetic variation in the glutinous sorghum germplasm in southwest China. Whole-genome sequences of ‘BTx623’ provide sequencing templates for genetic variations and molecular breeding. In this study, we compared whole genome resequencing data for six representative sorghum accessions with data for the BTx623 reference genome (Paterson et al., 2009; Zheng et al., 2011) to detect genomic variations and unique candidate genes for further study and to further characterize the elite sorghum germplasm of southwest China.

**MATERIALS AND METHODS**

**Plant Materials and Genome Sequencing**

Six S. bicolor accessions from southwest China were used in this study. Accession No.1 (Line 1) is a sorghum landrace called ‘Niu’ from the Guizhou province. Line 2 is the famous cultivar Hongyingzi, which is used to make Jiang-flavor liquor; it has the largest planting area in southwest China. Line 3 is a near isogenic line accession of Hongyingzi that has a similar phenotype as Line
2. Line 4 is the sorghum cultivar Hongzhenzhu, which can also be used to make Jiang-flavor liquor. Line 5 is the cultivar Qiango, from Sichuan province, which is used by several famous distilleries to produce Luzhou-flavor liquors (also called “Lu aroma”) liquor in China.

Seeds were planted in pots with a sand–peat mixture and maintained in a greenhouse. DNA was extracted from young freeze-dried leaf tissue via the cetyltrimethylammonium bromide method (Stein et al., 2001). Following quality assessments, genomic DNA was randomly sheared to ~350 bp in size with a Covaris ultrasonic processor (S220, Covaris, Woburn, MA). We used a TruSeq Library Construction Kit (Illumina Inc., San Diego, CA) to add an “A” base to the ends of DNA with double-strand breaks, followed by ligation of DNA adapters, product purification, and polymerase chain reaction amplification. The libraries were sequenced on the Illumina HiSeq 2000 platform (Illumina) by the Novogene Bioinformatics Institute, Beijing, China (Cheng et al., 2015). The BTx623 reference genome sequences were downloaded from the Ensemble database (ftp://ftp.ensemblgenomes.org/pub/plants/release-30/fasta/sorghum_bicolor/dna, accessed 26 July 2018).

Filtering Reads and Mapping Reads

Paired-end reads 150 bp long were determined and the clean reads were collected from sequenced reads, which were preprocessed to remove adaptors and low-quality paired reads. The following criteria were used to remove low-quality reads: (i) the presence of adapters, (ii) low-quality value (Phred score ≤ 5) for >50% of bases, and (iii) >10% “N”s. Alignment of the clean reads of each strain to the *S. bicolor* genomic reference was performed via the Burrows–Wheeler Aligner (Li and Durbin, 2009), and duplicated reads were removed and coverage values were calculated with SAMtools (Li et al., 2009).

Identification and Annotation of Variations

Single-nucleotide polymorphisms and indels (<50 bp) were identified via the MPILEUP tool in SAMtools (Cheng et al., 2015; Li et al., 2009). To reduce the SNP detection error rate, we filtered out the SNPs for which the supported read number was <4 or the quality value was <20. The bam file produced from the mapping procedure was analyzed for structural variations with BreakDancer with the default parameters (Chen et al., 2009). Single-nucleotide polymorphisms, indels, and structural variations were displayed by the Savant Genome Browser (Fiume et al., 2012). Copy number variations (CNVs) were detected with the CALL function in CNVnator (Abyzov et al., 2011). Functional annotation of all genetic variants was performed via ANNOVAR (Wang et al., 2010). Circos plots were created with Circos (Krzywinski et al., 2009).

Genetic relationships among these populations were analyzed via the unweighted pair group method with arithmetic mean method in PowerMarker (Liu and Muse, 2005) based on the Nei genetic distance (Nei, 1973). The reliability of the tree topology was tested by 1000 bootstrapped computations in PowerMarker and strict consensus maps were constructed with the Consense package in PHYLIP version 3.67 (Felsenstein, 1993).

Nucleotide diversity was calculated via Watterson’s θ (Watterson, 1975) and Tajima’s π (Tajima, 1989). Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis was performed with the online Database for Annotation, Visualization and Integrated Discovery (Huang et al., 2009). De novo assembly of unmapped reads was accomplished in the SOAPdenovo program (Luo et al., 2012).

Data Availability

Raw sequence data in fastq format have been deposited in the Short Read Archive database (https://www.ncbi.nlm.nih.gov/sra). The deposition number is SRP120997.

RESULTS

Resequencing yielded 1.68 billion paired end reads comprising 258.26 Gb of high-quality raw data. The average sequence depth was 54.3× coverage (range, 31.6–74.6×) for the six sorghum lines. The resequencing data were sufficient for genetic variation analysis (Fig. 1).
Table 1. Number and distribution of single-nucleotide polymorphisms in the six sorghum genomes.

<table>
<thead>
<tr>
<th>Code</th>
<th>Upstream</th>
<th>Stop gain</th>
<th>Stop loss</th>
<th>Synonymous</th>
<th>Nonsynonymous</th>
<th>Intronic</th>
<th>Downstream</th>
<th>Intergenic</th>
<th>Heterogenous rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.1</td>
<td>54,227</td>
<td>608</td>
<td>122</td>
<td>21,575</td>
<td>27,456</td>
<td>51,283</td>
<td>48,813</td>
<td>1,688,866</td>
<td>0.464</td>
</tr>
<tr>
<td>No.2</td>
<td>57,251</td>
<td>655</td>
<td>131</td>
<td>22,776</td>
<td>29,203</td>
<td>54,244</td>
<td>51,608</td>
<td>1,747,087</td>
<td>0.673</td>
</tr>
<tr>
<td>No.3</td>
<td>56,693</td>
<td>645</td>
<td>118</td>
<td>22,584</td>
<td>29,033</td>
<td>53,602</td>
<td>51,049</td>
<td>1,734,838</td>
<td>0.625</td>
</tr>
<tr>
<td>No.4</td>
<td>55,901</td>
<td>650</td>
<td>121</td>
<td>22,429</td>
<td>28,770</td>
<td>53,276</td>
<td>50,400</td>
<td>1,718,004</td>
<td>0.573</td>
</tr>
<tr>
<td>No.5</td>
<td>52,947</td>
<td>614</td>
<td>122</td>
<td>21,597</td>
<td>27,458</td>
<td>51,240</td>
<td>48,107</td>
<td>1,662,998</td>
<td>0.524</td>
</tr>
<tr>
<td>No.6</td>
<td>52,129</td>
<td>604</td>
<td>116</td>
<td>21,122</td>
<td>27,037</td>
<td>49,573</td>
<td>47,154</td>
<td>1,658,804</td>
<td>0.463</td>
</tr>
</tbody>
</table>

Single-Nucleotide Polymorphism Variations

We identified a large number of SNPs, indels, CNVs, and structural variations compared with the reference genome BTx623. In total, 2365,363 SNPs were identified, 63,446 of which were located in coding regions. Line 2 to Line 4 had higher levels of SNPs than the other three sorghum lines. The annotation results illustrated that Line 2 had the highest numbers of SNPs in upstream, downstream, exon, and intron regions. The SNP heterogeneity rate, which was based on the number of SNPs, exceeded 0.5 for Line 2 to Line 5, whereas Line 1 and Line 6 displayed relatively low heterogeneity (Table 1).

To obtain representative molecular markers, we filtered the SNPs, subsequently identifying 714,642 polymorphic SNP markers among the six sorghum lines (Fig. 2). Chromosome 5 had the highest number of SNPs (107,783), followed by chromosomes 10, 1, 3, 2, 4, 6, 8, and 7. Chromosome 9 had the fewest SNPs (44,442). Chromosomes 5 and 10 were relatively short (62.24 and 61.08 Mb, respectively) and thus had higher polymorphic molecular marker densities. We also combined these polymorphic SNP markers with markers from the sorghum genome SNP database (Luo et al., 2016) and consequently identified 441 common SNP markers among 54 sorghum accessions (Supplemental Table S1) with different geographic origins. The UPGMA trees were constructed by PowerMarker software (Fig. 3A). Four main clustered groups were obtained in this study. The first group included two Sorghum propinquum accessions and several Chinese sorghum lines, and the second group included two Sorghum margaritiferum accessions from Guinea in Africa. The other two groups revealed that some wild sorghum accessions and landraces from Ethiopia, Nigeria, and Kenya had more frequent SNPs and the variations of some current Chinese cultivars were similar to those of current American cultivars. Moreover, the six sorghum lines from southwest China formed an independent subgroup (Fig. 3A). After examination of a large number of SNP markers, the results illustrated that Line 2 and Line 3 formed an independent subgroup and had similar variation to Line 4, which was bred in the same geographic area. However, Line 5 and Line 6 exhibited different genetic variations from the other subgroup. The results illustrated that the six sorghum varieties had strong genetic relationships with their breeding environments.

We calculated the number of segregating sites per site (Watterson, 1975) within 1-Mb sliding windows across the genome to identify the regions with low sequence diversity in each chromosome (Lai et al., 2010). The data revealed that a bimodally distributed histogram contained window values of ~0.003397 and ~0.003422 SNPs per site (Supplemental Fig.S1). Two components of the SNP window classes fitted well with a normal mixture model. We calculated the probability that any sliding window belonged in a specific class (Lai et al., 2010), revealing that 14 genomic intervals had reduced numbers of SNPs (Supplemental Table S2). The average length of these low-diversity regions was 2.6 Mb and the longest regional length of 5.1 Mb was found on chromosome 1. In regions with reduced SNP density, we identified a number of genes that participated in glycolysis, amino acid metabolism, and phytohormone biosynthesis. Interestingly, one gene (Sb03g035250) had roles in pathways, including flavonoid, suberin, and phenylpropanoid biosynthesis. Sb02g011470 participated in phenylalanine biosynthesis, and Sb03g021040 controlled the conversion of 14-demethyl lanosterol to 4β-methylzymosterol-4α-carboxylate during steroid biosynthesis. The overall genome diversities among the six sorghum accessions were 0.000445 (Watterson’s θ) and 0.000466 (Tajima’s π), in line with some previous findings (Morris et al., 2013) but lower than those obtained from a large range of geographic populations (Mace et al., 2013). The results of Tajima’s D test revealed no obvious differences in the 10 chromosomes among the six sorghum lines. However, statistical significance was observed for some chromosomal regions, such as those of 60.7 to 61.3 Mb (P < 0.05), in which the plant maturity gene Ma was located (Mace and Jordan, 2010) (Fig. 4). These specific regions may reflect different selection pressures in the breeding process of glutinous sorghum from southwest China.

Genes with Indels and CNVs

We detected genome-wide indels 1 to 10 bp in size (Fig. 5A) and calculated the number of genes featuring these alterations (Fig. 5B). After we excluded single-base indels, 3- and 6-bp indels had relatively high frequencies in the gene regions. Moreover, the distributions of indel lengths in gene and intergenic regions were similar among the six sorghum lines. KEGG enrichment analysis revealed that indels were significantly (P < 0.05) enriched in genes involved in steroid, brassinosteroid, and secondary metabolite biosynthesis; pentose and glucuronate interconversions; metabolic pathways; basal transcription; active transport (i.e., ABC transporters); and inositol phosphate metabolism (Table 2).
Meanwhile, 2643 genes had 5407 CNVs among the six sorghum lines and 1226 genes had common CNVs. These genes encode proteins involved in oxidative phosphorylation; photosynthesis; steroid, starch, and sucrose metabolism; flavonoid biosynthesis; and ubiquitin-mediated proteolysis. Kyoto Encyclopedia of Genes and Genomes enrichment analysis demonstrated that CNVs were significantly \((P < 0.05)\) enriched in metabolic pathways, photosynthesis, oxidative phosphorylation, and homologous recombination.

**Genetic Variations among the Six Sorghum Lines**

We analyzed large-effect SNPs that may affect gene function. The study results identified 837 SNPs that are expected to induce premature stop codons or remove the annotated stop codons. The genes in which codon changes were induced were mainly involved in cutin and wax biosynthesis (Sb02g027540), amino acid and fructose metabolism (Sb03g009800 and Sb03g027065), plant hormone signal transduction, and ubiquitin-mediated proteolysis. To determine the unique large-effect variations of Line 2, we compared the large-effect SNPs and indels (nonsynonymous or frameshift variations) in the accession with...
those in Line 5 and Line 6. In total, 64 genes with large-effect SNPs (Supplemental Table S3) and 18 genes with large-effect indels (Supplemental Table S4) were identified (three genes featured both large-effect SNPs and indels). These genes were primarily involved in basic metabolic pathways, including purine metabolism, photosynthesis, and plant hormone signal transduction. Among them, two genes participated in linoleic acid (Sb01g045240) and sphingolipid metabolism (Sb03g030030).

Meanwhile, large genetic differences were noted between Chinese and foreign sorghum accessions. We filtered a total of 74.87 Mb of unmapped genome in this study (Supplemental Table S5) and then BLAST-searched these contigs in five pathway databases (Gene Ontology, InterPro, KEGG, Swiss-Prot, and TrEMBL) to annotate the functions of some high-confidence genes with homology support. Two hundred forty-one genes in Line 2 were found to be homologous with genes in crops, such as rice (Oryza sativa L.), maize (Zea mays L.), and barley (Hordeum vulgare L.), versus 140, 143, 152, 142, and 150 genes in Line 1, Line 3, Line 4, Line 5, and Line 6, respectively (Fig. 6). Most of the high-confidence genes participate in basic metabolism, coding heat-shock protein, or putative disease resistance. Sorghum variety

Fig. 4. Sequence diversity level, gene density, and gene location on chromosome 1.

Fig. 5. Distribution of insertion and deletion (indel) lengths in the sorghum genome. (A) Number of indels of different lengths in the whole genome. (B) Number of genes containing indels in the genome.
Table 2. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis of significantly ($P<0.05$) enriched insertions and deletions in six sorghum accessions.

<table>
<thead>
<tr>
<th>KEGG ID</th>
<th>Pathway name</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>sbi00100</td>
<td>Steroid biosynthesis</td>
<td>$2.00 \times 10^{-4}$</td>
</tr>
<tr>
<td>sbi00905</td>
<td>Brassinosteroid biosynthesis</td>
<td>$2.00 \times 10^{-4}$</td>
</tr>
<tr>
<td>sbi00404</td>
<td>Pentose and glucuronate interconversions</td>
<td>$4.60 \times 10^{-1}$</td>
</tr>
<tr>
<td>sbi01110</td>
<td>Biosynthesis of secondary metabolites</td>
<td>$1.30 \times 10^{-2}$</td>
</tr>
<tr>
<td>sbi01100</td>
<td>Metabolic pathways</td>
<td>$3.20 \times 10^{-2}$</td>
</tr>
<tr>
<td>sbi03022</td>
<td>Basal transcription factors</td>
<td>$2.80 \times 10^{-2}$</td>
</tr>
<tr>
<td>sbi02010</td>
<td>ABC transporters</td>
<td>$2.30 \times 10^{-2}$</td>
</tr>
<tr>
<td>sbi00562</td>
<td>Inositol phosphate metabolism</td>
<td>$3.30 \times 10^{-2}$</td>
</tr>
</tbody>
</table>

Line 2 also had the largest number of unique homologous genes (104), whereas the remaining accessions had similar unmapped sequence variations.

**DISCUSSION**

The development of sequencing technologies has permitted studies of plant genetic variations, molecular domestication, and genome-wide associations. The BTx623 genome provides a useful reference for genome studies (Paterson et al., 2009), as a large number of SNPs and functional genes have been identified in different geographic populations in this genome (Mace et al., 2013; Morris et al., 2013). In this study, we identified 2.37 million SNPs and 0.39 million indels among the six sorghum lines from southwest China.

The UPGMA trees generated in this study revealed that sorghum accessions originating from African regions were distributed into several subgroups and they exhibited high levels of genetic variation. The northeast quadrant of Africa is a sorghum domestication center, and *S. margaritferum* had an intermediate phenotype between wild (*S. bicolor* ssp. *verticilliflorum*) and cultivated sorghum (Grenier et al., 2004; Mace et al., 2013; Morris et al., 2013). The two accessions of *S. propinquum* (369–1 and 369–2) formed an independent subgroup in this study, indicating that this species is divergent from Asian *S. bicolor* accessions in both genotype and phenotype (Mace et al., 2013) and is a potential resource for comparative genomic studies of *S. bicolor*. Two hypotheses regarding the origins of Chinese sorghum have been proposed. One hypothesis proposes that the crop originated in India and Africa (Grenier et al., 2004), and the other postulates that it has multiple origins, including China and other geographic regions (Grenier et al., 2004; Qingshan and Dahlberg, 2001). In this study, most Chinese sorghum varieties formed a distinct subgroup and exhibited genetic similarity to *S. propinquum* (Fig. 2A), supporting the Asian origin of the crop. The six sorghum accessions from southwest China formed an independent subgroup, although detailed clustering could not be performed because of the limited number of common SNP markers. Cluster analysis of our six sorghum lines indicated that Line 2 was an improved variety that formed a subgroup with its near isogenic line, Line 3. Southwest China includes the Guizhou, Sichuan, and Yunnan provinces, which feature multiple mountain environments, and conventional sorghums in this region formed an elite germplasm during the breeding process and its application through both natural and artificial selection. This special germplasm will be a potential resource for sorghum breeding and improvement in China.

The overall genomic diversity of the six lines was relatively low and no obvious difference was observed.

---

**Fig. 6.** Numbers of homologous genes identified via unmapped analysis. Red, green, blue, yellow, and pink represent sorghum accessions Nos 1, 2, 4, 5, and 6, respectively.
between Watterson’s θ and Tajima’s π for most chromosomal regions. Thus the conventional sorghum lines have a close genetic relationship, possibly because all six lines were bred from landraces in the same geographic region. However, Tajima’s D test revealed significant diversity in some chromosomal regions. In a previous study on differences between Watterson’s θ and Tajima’s π (Mace and Jordan, 2010), significant differences were noted for four of the eight examined genes. Maₚ, which controls plant maturity, displayed differences in both genotype and phenotype among the six sorghum lines (Line 2 had a growth period of 120–130 d, Line 5 had a growth period of ~110 d, and Line 6 had a growth period of ~130 d). Regional differences in chromosomal diversities may be evidence of selection pressure in the process of germplasm domestication, and further research is needed. Meanwhile, genomic intervals with a reduced number of SNPs mainly contained genes involved in basic metabolism and abiotic and biotic stress tolerance, consistent with the results in previous reports (Winkel-Shirley, 2001; Zheng et al., 2011). This finding provides further evidence that some chromosomal regions experienced obvious selection pressure during germplasm formation, consistent with the findings in maize (Lai et al., 2010). Because of the small sample sizes and the limitations of the statistical methods, some candidate genes could not be detected in this study.

In addition, sorghum lines in this study had higher numbers of genetic variations when we considered indels and CNVs (Zheng et al., 2011). The large number of genes with 3- or 6-bp indels illustrated that the genome preferred insertions or deletions of one or two codons to avoid frameshift variations during domestication. A previous report illustrated that indels with lengths that were not a multiple of 3 bp were uncommon in coding regions, as these mutations may be harmful to sorghum survival (Zheng et al., 2011). For some gene families, the indels and CNVs may display similar survival and distribution patterns. It has been reported that CNVs may contribute to heterosis in crop domestication and stress (or disease) tolerance (Springer et al., 2009). Meanwhile, the KEGG enrichment results illustrated that indels and CNVs occurred in genes involved in a number of basic metabolic pathways as well as steroid biosynthesis and photosynthesis, and these changes could explain some important phenotypic variations in the Chinese sorghum population.

Some of the identified genes with large-effect SNPs are involved in cutin and wax biosynthesis or plant hormone signal transduction. In line with this finding, these six sorghum lines have higher than average plant heights (2–2.5 m) and biomass. For Line 2, 79 genes affected by large-effect SNPs or frameshift indels were identified, and these genes may play an important role in the elite traits of this special variety. Among the six sorghum lines (Fig. 6), compared with the reference genome, some unique homologous genes were identified, which could be used to study the origins or breeding selection of the Chinese sorghum germplasm. We identified 78 common high-confidence genes among the six landraces, and Line 2 had 104 unique homologous genes. According to the functional genome, Line 2 acquired some special genetic variations during breeding.

CONCLUSIONS

We performed whole-genome analysis of SNPs, indels, and CNVs in six sorghum lines from southwest China. The unique variations and genes detected in this study will serve as a foundation for studying genetic variations and germplasm domestication in Chinese sorghum populations. Moreover, the SNP and indel markers discovered in this study will provide useful tools for identifying elite alleles and remarkable traits. These findings will be helpful for improving sorghum and other crops.

Supplemental Information

Supplemental Table S1. Category and origins of 54 sorghum accessions.
Supplemental Table S2. Genomic intervals of reduced numbers of single-nucleotide polymorphisms.
Supplemental Table S3. Sixty-four genes containing large-effect single-nucleotide polymorphisms in sorghum variety No. 2.
Supplemental Table S4. Eighteen genes in sorghum variety No. 2 containing large-effect indels.
Supplemental Table S5. Number and length of unmapped contigs in six sorghum genomes.
Supplemental Fig. S1. Distribution patterns of genome-wide single-nucleotide polymorphisms in 1-Mb sliding windows from 10 chromosomes.

ACKNOWLEDGMENTS

This project was supported by the Foundation of Moutai Group post-doctoral program from China Kweichow Moutai Distillery Co., Ltd. (2015005). All funders had no role in the study design, data collection and analysis, the decision to publish, or preparation of the manuscript. We thank Mr. Youneng Tu (Hongyingzi Agricultural Development of Science and Technology Co., Ltd., China) for his kind help in providing sorghum accessions Line 1 to Line 4, and we are grateful to China Kweichow Moutai Distillery Co., Ltd. for funding this study. All authors have approved the manuscript.

REFERENCES


Zhang, D., J. Li, R.O. Compton, J. Robertson, V.H. Goff, E. Epps et al. 2015. Comparative genetics of seed size traits in divergent cereal lineages represented by Sorghum (Panicoideae) and Rice (Oryzoidae). G3: Genomes Genet. 5: 1117–1128. doi:10.1534/g3.115.017590.
