Characterization of RFLP probe sequences for gene discovery and SSR development in *Sorghum bicolor* (L.) Moench

S. J. Schloss¹, S. E. Mitchell², G. M. White³, R. Kukatla², J. E. Bowers⁴, A. H. Paterson⁵ and S. Kresovich⁶

Communicated by D. Hoisington

**Abstract.** In this study, we collected and analyzed DNA sequence data for 789 previously mapped RFLP probes from *Sorghum bicolor* (L.) Moench. DNA sequences, comprising 894 non-redundant contigs and end sequences, were searched against three GenBank databases, nucleotide (nt), protein (nr) and EST (dbEST), using BLAST algorithms. Matching ESTs were also searched against nt and nr. Translated DNA sequences were then searched against the conserved domain database (CDD) to determine if functional domains/motifs were congruent with the proteins identified in previous searches. More than half (500/894 or 56%) of the query sequences had significant matches in at least one of the GenBank searches. Overall, proteins identified for 148 sequences (17%) were consistent among all searches, of which 66 sequences (7%) contained congruent coding domains. The RFLP probe sequences were also evaluated for the presence of simple sequence repeats (SSRs) and 60 SSRs were developed and assayed in an array of sorghum germplasm comprising inbreds, landraces and wild relatives. Overall, these SSR loci had lower levels of polymorphism ($D = 0.46$, averaged over 51 polymorphic loci) compared with sorghum SSRs that were isolated by library hybridization screens ($D = 0.69$, averaged over 38 polymorphic loci). This result was probably due to the relatively small proportion of di-nucleotide repeat-containing markers (42% of the total SSR loci) obtained from the DNA sequence data. These di-nucleotidic markers also contained shorter repeat motifs than those isolated from genomic libraries. Based on BLAST results, 24 SSRs (40%) were located within, or near, previously annotated or hypothetical genes. We determined the location of 19 of these SSRs relative to putative coding regions. In general, SSRs located in coding regions were less polymorphic ($D = 0.07$, averaged over three loci) than those from gene flanking regions, UTRs and introns ($D = 0.49$, averaged over 16 loci). The sequence information and SSR loci generated through this study will be valuable for application to sorghum genetics and improvement, including gene discovery, marker-assisted selection, diversity and pedigree analyses, comparative mapping and evolutionary genetic studies.

**Keywords.** Database search - BLAST - EST - Microsatellite - Germplasm - Genetic diversity
Introduction

Sorghum [Sorghum bicolor (L.) Moench], a grain crop originating in Africa, is grown worldwide for both food and forage (Doggett 1988). Sorghum is a diverse genus consisting of both cultivated and wild species. The most important agronomic form is S. bicolor ssp. bicolor (2n = 20), a largely self-pollinated diploid comprising five cultivated races (bicolor, caudatum, durra, guinea and kafir) and their hybrids (Harlan and de Wet 1972).

Sorghum and maize (Zea mays L.) shared a common ancestor as recently as 20-24 million years ago (Gaut and Doebley 1997), while the lineages leading to rice (Oryza sativa L.) and sorghum/maize diverged slightly more than 50 million years ago (Chen et al. 1998). Because of its evolutionary history and intermediate genome size (approximately 690 Mb), sorghum may provide the appropriate link for extending genetic information derived from the small-genome model grass, rice (440 Mb), to large-genome grasses such as maize (2,500 Mb) (Arunuganathan and Earle 1991). To become a bridge for grass genomics, however, detailed genetic and physical maps for sorghum must first be developed. Dense genetic maps are available for both rice (Harushima et al. 1998) and maize (Davis et al. 1999). Although recent maps show improvement, the sorghum genetic maps are less saturated (Ming et al. 1998; Peng et al. 1999; Bhattramakki et al. 2000). Current efforts, therefore, have focused both on developing and mapping new molecular markers in sorghum and on deriving DNA sequence information from existing markers (Bowers et al. 2000; Ventelon et al. 2001).

In this study, DNA sequences were obtained from mapped sorghum RFLP probes. Specific objectives were to identify probes that most likely contain sorghum genes by searching the sequences obtained against the public DNA sequence databases and, when possible, to convert the RFLP probes to simple sequence repeat (SSR) markers. SSR loci are highly polymorphic among natural plant populations and inbred lines (Innan et al. 1997; Senior et al. 1998) and, compared to RFLPs, these PCR-based markers are more easily assayed. The DNA sequence data, putative gene identities, EST annotations and SSR markers presented here will be valuable for future work in sorghum marker-assisted selection, comparative grass mapping projects, and population and evolutionary genetic studies.

Materials and methods

DNA sequencing and preliminary analysis

The genomic RFLP probes (n = 789) were obtained from the University of Georgia, Center for Applied Genetic Technologies (Athens, Ga.). These probes were mapped
previously in a *S. bicolor × Sorghum propinquum* F2 population (Chittenden et al. 1994; Bowers et al. 2000). Clone insert sizes ranged from approximately 350 bp to 1,500 bp. All probes were sequenced twice from both directions with BigDye-terminator sequencing chemistry (Applied Biosystems) and sequence data were collected on an automated DNA sequencer (Applied Biosystems, Model 377). DNA sequences were edited, aligned, and checked for redundancies using Sequencher (Gene Codes Corporation). These sequences were deposited in the GenBank Genome Survey Sequences database (dbGSS) under accession numbers BH245205-BH246341.

**Database searches**

The GenBank databases were accessed through the National Center for Biotechnology Information, National Library of Medicine (Bethesda, Md.) (http://www.ncbi.nlm.nih.gov). Gapped BLASTN and BLASTX algorithms (Altschul et al. 1997) were used to search the nucleotide (nt), protein (nr) and EST (dbEST) databases. The conserved domain database (CDD; complete database including Smart v. 3.3 and Pfam v. 6.5) was searched using Reverse Position Specific (RPS)-BLAST. RPS-BLAST compares a protein query sequence to a position-specific score matrix prepared from the underlying conserved protein domain alignment (Altschul et al. 1997). For this search, DNA sequences were translated in six reading frames with Transeq (http://www.sander.embl-ebi.ac.uk/Services/emboss/transeq.html) using the standard genetic code.

RFLP probe sequences were searched against the nt and nr databases using BLASTN and BLASTX, respectively. DNA sequences were then searched against dbEST using BLASTN, and the full-length EST sequences identified by this search were queried against nt and nr. The nt, nr and EST databases were downloaded on June 22, 2000. Searches of the CDD were performed on-line (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) on December 10, 2000.

BLAST searches were performed with default parameters except for the initial $E$-value, which was set at 10. Output was limited to the top ten sequence alignments (matches) per query based on the score. Matches with $E$-values $\leq 1 \times 10^{-8}$ were selected for further evaluation, those with $E$-values between $1 \times 10^{-8}$ and $1 \times 10^{-11}$ were inspected individually, and alignments shorter than 50 bp were discarded. Since multiple databases were searched, we used an $E$-value threshold that was more stringent than recommended for BLAST searches against single databases ($E$-value $\leq 0.005$) (Anderson and Brass 1998). When a DNA sequence produced significant matches in multiple searches, annotations were compared to confirm that results from each search were consistent (i.e., the same protein was identified). For the CDD search, all default parameters were employed.

**Development of SSRs**
Identification of repeat motifs

Sorghum RFLP probe sequences were searched for all possible 2-6 base-pair repeat motifs using a program developed by Sam Cartinhour (USDA-ARS Center for Agricultural Bioinformatics). Di-, tri-, and longer motifs (tetra-, penta- and hexa- nucleotides) containing six, five and four repeat units, respectively, were identified. Optimal primer sequences for amplifying each SSR locus were obtained with Primer 0.5 (Daly et al. 1991) (http://www-genome.wi.mit.edu/ftp/pub/software/primer.0.5). Primer pairs were synthesized commercially. For fluorescence-based detection, the 5′ end of the "forward" primer of each SSR locus was labeled with either FAM, TET or HEX dyes (Applied Biosystems).

Plant material and DNA extraction

SSR markers were initially evaluated in a panel of 25 sorghum DNAs comprising 22 inbred lines (16 proprietary lines from Pioneer Hi-Bred International and six public accessions) and three landraces from the International Crops Research Institute for the Semi-arid Tropics (ICRISAT) (Table 1, nos. 1-25). The proprietary accessions included one set of parental and hybrid lines for evaluating the inheritance (i.e., codominance) of SSRs. Based on polymorphism levels and compatibility of fluorescent dye labels for multiplexing samples, a subset of 33 SSRs was selected and assayed in 12 additional public accessions representing geographically diverse samples of the five sorghum races and wild material (Table 1, nos. 26-37). The panel of germplasm used for the SSR assays was selected to represent: (1) a closely related pool of elite germplasm important to sorghum improvement in the U.S., and (2) broad geographic and racial representation of sorghum diversity. With this set of divergent reference pools, our goal was to establish the discriminatory power of the SSRs in closely related materials, as well as establish their usefulness across the range of sorghum.

DNA was isolated from individual plants (7-10-day old seedlings) following a standard CTAB extraction protocol (Doyle and Doyle 1987). The crude nucleic acid preparations were precipitated in 30% isopropanol, pellets were rinsed with 70% ethanol, dried, and suspended in 1X TE (10 mM Tris-HCl; 1 mM EDTA, pH 8.0).

PCR amplification and gel electrophoresis

Polymerase chain reactions (PCRs), sample preparation, electrophoresis, fluorescence-based detection and automated fragment sizing followed Matsuoka et al. (2002). To assure precision and reproducibility, PCR reactions and gel runs were replicated.

Diversity estimates and significance test

For each SSR locus, two measures of genetic diversity were calculated; the number of alleles per locus, and the gene diversity index (D) (Nei 1973). D was estimated as follows:
\[ D = \frac{n}{(n-1)} \left[ 1 - \sum (p_i^2) \right], \]

where \( n \) is the sample size and \( p_i \) is the relative frequency of the \( i \)th allele.

To test the null hypothesis that inbreds and diverse accessions belong to the same population based on their allele frequencies, a Monte Carlo estimation for the Pearson chi-square test (10,000 samples) was performed using the FREQ procedure in SAS (v. 8.2; SAS Institute, Cary, N.C.).

---

**Results**

**Characterization of RFLP probe sequences**

DNA sequences were obtained from 789 RFLP probes. Approximately 60% of the edited forward and reverse sequences from the same probe overlapped. The resulting contigs ranged from 195 bp to 1,165 bp with an average length of 730 bp. Non-overlapping sequences from probes with larger inserts (40% of the probes sequenced) were truncated at the first ambiguous base and averaged 450 bp in length. Redundant sequences were encountered for 30% of the probes (224 redundant probes comprising 104 unique sequences). Probes with redundant sequences generally mapped to the same chromosomal location (Bowers et al. 2000) and, in these cases, one sequence representative was used for BLAST searches and SSR screening. Redundant groups of probes that did not co-map were excluded from further analysis. In total, 894 query sequences were obtained, including both contigs and non-overlapping end sequences from 639 non-redundant RFLP probes.

**Database searches**

Fifty six percent (500/894) of the query sequences had significant alignments in at least one database (nt, nr or dbEST), and 31% of these (155/500) were significant in all three searches. Conversely, there were 118 DNA sequences with alignments unique to nt/nr, and 120 that matched ESTs only. For RFLP probes with significant similarity to ESTs, the corresponding full-length EST sequences were extracted and queried against nt and nr. Two hundred and twenty three ESTs had significant matches in nt/nr, and 148 of these annotations were consistent with results from all BLAST searches. Results from searching the CDD database for conserved domains characteristic of specific protein families provided further validation of these sequence annotations. Conserved motifs (Table 2) consistent with previous annotations were identified in 66 DNA sequences (45% of the 148 queries).

**Proteins identified**
Because of space constraints, complete listings of BLAST results from nt, nr, dbEST and CDD searches are not presented here. This information is available at the Cornell University, Institute for Genomic Diversity website (http://genotype.igd.cornell.edu/myapp/servlets/AnnotTable). BLAST results and domain annotations for RFLP probes with conserved protein domains, however, are shown in Table 2. For all searches combined, 52% (263/500) of matches were to "unknown" proteins. This result is most likely due to the large number of unannotated "hypothetical" or "putative" genes predicted by ORF-finding softwares that are present in the public databases. Known proteins were identified in 42% of the total matches (208/500), and two-thirds of these proteins represented plant genes. Transposable elements (either transposon-specific proteins or long terminal repeats) were identified in 6% of all significant matches (29/500). The small number of transposable elements identified is not surprising because the RFLP probes were selected low-copy sequences. As might be expected, a higher proportion of the 148 probe sequences that were consistently annotated by all searches, including those done using the full-length ESTs as queries, matched known genes compared to sequences with significant matches in one search only.

**Identification and characterization of SSRs**

One hundred SSR loci were identified from 894 DNA sequences, and primer pairs were successfully designed for 74 SSR markers from 69 RFLP probes. Based on preliminary genotyping results, 14 potential SSR markers were discarded, primarily due to failed or erratic amplification. The remaining 60 markers were used to evaluate levels of genetic diversity in sorghum accessions (Table 1, nos. 1-25).

**Genetic diversity analysis**

Primer sequences, diversity measures, and other relevant information for the 60 SSR loci are presented in Table 3. Eighty five percent (51/60) of SSR loci were polymorphic in the initial screening of accessions (Table 1, nos. 1-25). Polymorphic loci averaged 3.4 alleles per locus with an average diversity index ($D_{avg}$) of 0.46. Twenty eight SSR markers were variable among the parental lines and hybrid, and all of these exhibited single-locus, codominant inheritance. As would be expected, SSR loci containing di-nucleotide repeats were the most abundant and polymorphic marker type; 24 of 25 loci with di-nucleotide repeats were polymorphic with $D_{avg} = 0.49$. Although a smaller proportion (27/35) of SSRs with longer repeat motifs were polymorphic, these markers were nearly as informative ($D_{avg} = 0.43$) as the di-nucleotide markers. In general, loci with tetra-nucleotide repeats were slightly more polymorphic than the tri-nucleotide repeat-containing loci (3.2 alleles/locus with $D_{avg} = 0.48$ compared to 2.7 alleles/locus with $D_{avg} = 0.42$, respectively).

For 24 loci (47% of polymorphic SSRs) we observed at least one allele that did not show step-wise variation in size (i.e., size differences among some alleles were not a multiple of the SSR core repeat unit) (Table 3). In general, alleles from loci with di- and tri-nucleotide repeats exhibited step-wise allele size distribution more frequently than
markers with longer repeats (65% of the di- and tri-nucleotide repeat-containing SSRs compared to 32% of loci with tetra-, penta- and hexa-nucleotide repeats).

A subset of 33 polymorphic SSRs was assayed in a larger population that included the initial 25 lines (Table 1, nos. 1-25) and 12 additional accessions comprising geographically diverse material (Table 1, nos. 26-37). Data from four loci (Xcup24, Xcup48, Xcup64, Xcup67) were discarded, primarily because of poor amplification in the diverse material (>10% null alleles). Data obtained from the remaining SSR loci were used to estimate levels of genetic diversity for inbred lines (n = 22) and geographically diverse sorghum accessions (n = 15) (Table 4). Although the diverse material exhibited more variation at 18 of the 29 SSRs assayed, the overall amount of genetic diversity present in these accessions and the inbred lines was similar (Davg = 0.60 and 0.54; mean number of alleles per locus = 4.9 and 3.6, respectively). For most SSR loci, the loss of rare alleles in the inbreds was responsible for the generally lower D values observed in this group (data not shown). We should note that the elite sorghum inbred lines assayed in this study represent germplasm that is routinely used in sorghum breeding for hybrid development in the U.S. As such, these lines should encompass a relatively broad array of germplasm diversity.

Monte Carlo estimates of the exact p-values (<0.0001) for the Pearson chi-square test indicated that the inbred and diverse groups differed significantly in allele frequency at one or more of the 29 SSR loci evaluated. Because of substantial differences in allele frequencies, the D values of five SSR loci (Xcup06, Xcup13, Xcup32, Xcup33 and Xcup47) were at least 10% greater in the inbreds compared to the diverse accessions (Table 4). This result might be due to selection at these or other closely linked loci, or genetic drift, or it could be an artifact of small sample sizes.

SSRs in putative gene regions

Putative genes identified by BLAST searches and positions of SSRs relative to these coding regions are presented in Table 5. Twenty four of the 60 SSR markers used in the diversity analyses were located within, or near, previously annotated or hypothetical genes. These genes had various functions, including stress response, developmental regulation, cellular transport or metabolism. We were able to determine the locations of 19 SSRs relative to putative coding regions. Only three SSRs were located within protein coding regions. Two of these, Xcup34 and Xcup65, were invariant in the germplasm tested and the information content of the third locus, Xcup42, was low (D = 0.22). In contrast, all but one of the SSRs located in gene flanking regions, UTRs, or introns were polymorphic (for 20 polymorphic loci, Davg = 0.46).

Discussion

In this study, DNA sequences from mapped sorghum RFLP probes were analyzed for gene content, and repeat motifs were identified for development of SSR markers. To
maximize the amount of information captured, searches of multiple GenBank databases (nt, nr and dbEST) were performed. Significant matches from dbEST verified that the sequences were transcribed, and a final search of nt and nr using full-length ESTs confirmed the initial results. The cross-referencing of results from multiple similarity searches provides a means to validate the consistency of significant matches. Failure to meet the significance criteria, however, does not necessarily mean that a particular match provides no information. Considering that the RFLP probes analyzed in this study had relatively short inserts (730 bp, on average), it is likely that short regions of sequence similarity from legitimate coding regions were below the significance thresholds.

Three years ago, GenBank searches were an inefficient method for identifying genes in plant DNA sequences. For example, a search of the nt database resulted in significant matches for only 9% of the 259 barley RFLP probes assayed (Michalek et al. 1999). In the present study, putative genes (proteins or coding regions) were identified in 56% of sorghum queries. This increase in efficiency of gene discovery is most likely due to growth in the public databases. Over the past 2 years, the number of plant DNA sequences deposited in GenBank has increased exponentially, primarily due to the efforts of genome sequencing initiatives for Arabidopsis and rice, and large EST sequencing projects for a variety of plant species, including sorghum (http://www.ncbi.nlm.nih.gov/Database/index.html). Although similarity searches are becoming more productive for identifying plant genes, they are not the ultimate test for biological significance or gene function. What bioinformatics-based approaches do provide, however, is a means for homing in on particular DNA sequences that may play important roles in systems of interest to plant scientists and breeders.

Analysis of the RFLP probe sequences yielded 60 new SSR loci, 51 of which were polymorphic in an array of sorghum germplasm. In general, the genetic variation detected by these SSRs was lower than diversity estimates for SSR loci isolated from traditional genomic library screens. Diversity estimates (51 loci, 3.4 alleles/locus, $D_{avg} = 0.46$) were similar to values reported by Brown et al. (1996) (17 loci, 3.8 alleles/locus, $D_{avg} = 0.54$), but the sequence-derived SSRs were less polymorphic than markers developed by Kong et al. (2000) (38 loci, 4.8 alleles/locus, $D_{avg} = 0.69$). This observation might be due to the sorghum germplasm tested. More likely, however, the lower information content of sequence-derived SSRs was related to differences in the proportion of di-nucleotide repeat markers tested and disparities in repeat unit length. For the SSRs described here, slightly less than half the SSR markers contained di-nucleotide repeats (with nine repeat units/locus, on average), while 87% of SSRs isolated by Kong et al. (2000) were di-nucleotides (averaging 22 repeat units/locus). In general, SSRs with di-nucleotide repeats are the most polymorphic marker class, and a direct relationship exists between marker information content and the number of repeat units (Weber 1990; Innan et al. 1997; Schug et al. 1998). Traditional hybridization screens of small insert genomic DNA libraries usually target di-nucleotide repeat-containing SSRs because these sequences are highly represented in the genome (Condit and Hubbel 1991). Because SSRs containing longer repeat motifs (≥ 4 bp) occur less frequently, they are rarely isolated through standard library screening methods. SSRs derived from DNA sequence data have no such bias. Thirty seven percent of the SSR loci identified here contained either tetra-, penta- or
hexa-nucleotide repeats, and these markers were almost as informative as the di-
nucleotide SSRs ($D_{\text{avg}} = 0.43$ compared to $D_{\text{avg}} = 0.49$, respectively).

SSRs are hypothesized to mutate in multiples of the repeat unit, due to polymerase
slippage during DNA replication (Levinson and Gutman 1987). Here, we observed 24
SSRs, 47% of polymorphic loci, with allele size distributions that did not conform to this
simple model. We also found that SSRs with repeat motifs $\geq$ four nucleotides were twice
as likely to violate this model than loci with di- or tri-nucleotide repeats. A recent study
in maize has shown that length polymorphisms in 87% of the SSRs evaluated (mainly tri-
and tetra-nucleotide-containing loci) were probably due to indels in DNA flanking the
repeat motif and not to variation in repeat number (Matsuoka et al. 2002). Although the
occurrence of these complex mutation patterns for sorghum SSRs did not appear to be as
frequent as in maize, estimation of population parameters based on step-wise models
(Kimura and Crow 1964) should be used with the knowledge that the model may be
imprecise for describing allelic distributions at some SSR loci. Because technological
advances in allele sizing now permit routine discrimination of alleles that were once
overlooked (i.e., those that differ in length by one nucleotide), further studies and
discussions of how robust the models are to violation of the assumptions are needed.

Development of SSR markers, either by traditional library screening methods or library
enrichment, is laborious and expensive. Conversely, technological advances in
sequencing chemistry, instrumentation, throughput and data handling have significantly
reduced the costs of collecting and analyzing DNA sequence data. Sequence-based
approaches for developing molecular markers such as SSRs, therefore, have become both
operationally and economically feasible. Furthermore, DNA sequences derived from
mapped sorghum RFLP probes are valuable, not only for assaying genetic diversity
within germplasm collections and wild populations, but also for linking genetic and
physical maps among the grasses, and for cross-species gene discovery and genome
characterization.

Acknowledgements. We thank Sam Cartinhour (USDA-ARS, Center for Bioinformatics
and Comparative Genomics, Cornell University) for providing programs for identifying
simple sequence repeat motifs. Sam Cartinhour and David Schneider (Cornell University,
Theory and Simulation Science and Engineering Center) also helped with designing
BLAST search strategies. Proprietary sorghum inbred lines were kindly supplied by
Yilma Kebede (Pioneer Hi-Bred International, Johnston, Iowa). Julie Ho (Cornell
University, Institute for Genomic Diversity) provided advice and help with the statistical
analysis. We extend special thanks to Julie Ho and Joanne Labate (USDA-ARS, Plant
Genetic Resources Unit, Geneva, N.Y.) for a critical review of the manuscript. This
research was funded by National Science Foundation grant DBI-9872649 and Pioneer Hi-
Bred International, and S.J.S. was partially funded by the Lavallard Fellowship. This
work was submitted by S.J.S. to the Cornell University Graduate School in partial
fulfillment of the Master of Science degree.