Theor Appl Genet (2002) 105:912–920
DOI 10.1007/s00122-002-0991-4

S. J. Schloss · S. E. Mitchell · G. M. White
R. Kukatla · J. E. Bowers · A. H. Paterson
S. Kresovich

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

Received: 24 August 2001 / Accepted: 17 January 2002 / Published online: 30 July 2002

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-nucleotide 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

Communicated by D. Hoisington

S.J. Schloss
Department of Plant Breeding and Institute for Genomic Diversity, Cornell University, 157 Biotechnology Building, Ithaca, NY 14853, USA

S.E. Mitchell · R. Kukatla
Institute for Genomic Diversity, Cornell University, 151 Biotechnology Building, Ithaca, NY 14853, USA

G.M. White
Institute for Genomic Diversity, Cornell University, 157 Biotechnology Building, Ithaca, NY 14853, USA

J.E. Bowers
Center for Applied Genetic Technologies, The University of Georgia, Athens, GA 30602, USA

A.H. Paterson
Department of Crop Sciences and Center for Applied Genetic Technologies, The University of Georgia, Athens, GA 30602, USA

S. Kresovich (✉)
Department of Plant Breeding and Institute for Genomic Diversity, Cornell University, 158 Biotechnology Building, Ithaca, NY 14853, USA
e-mail: sk20@cornell.edu
Tel.: +1-607-255-1492, Fax: +1-607-255-6249

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) (Arumuganathan 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* F_2 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 ≤ 1×10^{-8} were selected for further evaluation, those with E-values between 1×10^{-8} and 1×10^{-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 ≤ 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.

Table 1 Sorghum accessions analyzed

No.	Common name	IS No.[a]	PI No.[b]	Material	Species	Subspecies	Race	Origin
1–16	01–16	None	None	Inbred	*bicolor*	*bicolor*	NA[c]	Proprietary
17	BTx623	None	PI564163	Inbred	*bicolor*	*bicolor*	NA	USA
18	BTx3197	None	None	Inbred	*bicolor*	*bicolor*	NA	USA
19	M1 (SC566)[d]	IS7254C	PI533871	Inbred	*bicolor*	*bicolor*	guinea	Nigeria
20	Msumbji (SC283)	IS7173C	PI533869	Inbred	*bicolor*	*bicolor*	guinea	Tanzania
21	BR007 (101)	IS2749	PI267432	Inbred	*bicolor*	*bicolor*	bicolor	India
22	2031T11 (SC689)	IS2729C	PI533969	Inbred	*bicolor*	*bicolor*	caudatum	Uganda
23	P3730	IS2377	PI229835	Landrace	*bicolor*	*bicolor*	kafir-durra	S. Africa
24	Ramkel	IS1029	PI286232	Landrace	*bicolor*	*bicolor*	kafir	India
25	Bank oumano ziamri fing	IS3817	NSL51030	Landrace	*bicolor*	*bicolor*	guinea	Mali
26	Chinese Amber	IS12711	PI22913	Landrace	*bicolor*	*bicolor*	bicolor	China
27	65I2013 (G22)	IS2668	NSL51249	Landrace	*bicolor*	*bicolor*	bicolor	Uganda
28	Msumbji (65I1634)	IS7173	NSL50876	Landrace	*bicolor*	*bicolor*	guinea	Tanzania
29	Kokla (MN833)	IS12570	PI152705	Landrace	*bicolor*	*bicolor*	caudatum	Sudan
30	No.1 Gambela	IS12608	PI257595	Landrace	*bicolor*	*bicolor*	caudatum	Ethiopia
31	R1 (46)	IS2694	PI267380	Landrace	*bicolor*	*bicolor*	kafir-bicolor	Zimbabwe
32	A-5677	IS12687	PI302105	Landrace	*bicolor*	*bicolor*	bicolor	Ethiopia
33	178	IS12693	PI225905	Landrace	*bicolor*	*bicolor*	hybrid	Zambia
34	Unknown	IS3106	PI213900	Landrace	*bicolor*	*bicolor*	bicolor	Kenya
35	A-7171	None	PI302233	Unknown	*bicolor*	*verticilliflorum*	virgatum	Fmr. Sov. Union
36	Unknown	None	PI199869	Unknown	*bicolor*	*drummondii*	Unknown	S. Africa
37	Zhuronskiva	None	PI539065	Wild species	*halepense*	Unknown	Unknown	Kazakhstan

[a] International Crop Research Institute for the Semi-Arid Tropics (ICRISAT) identifier
[b] Plant Introduction; U.S. collection identifier
[c] Information not available
[d] Synonymous names in parentheses

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:

$$D = [n/(n-1)]\left[1 - \sum (p_i^2)\right],$$

where n is the sample size and p_i is the relative frequency of the ith 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)

Table 2 BLAST results and annotations for RFLP probes with conserved protein domains

Probe name	LG[a]	EST ID	nt/nr ID	Domain annotation from CDD	E-value[b]
pSB1311	B,E,I	AU100690	AAD37023	ABC transporter; elongation factor	3.00E^{-07}
pSB1249	H	AW285384	AAB67883	Acyl-CoA oxidase	5.00E^{-03}
pSB0705	Unmapped	BE494998	S57614	Adh-short; short chain dehydrogenase	8.00E^{-06}
pSB1028	B	AW676954	AAD03380	Adh-short; short chain dehydrogenase	4.00E^{-11}
pSB1801	B	C27369	AAD55139	2-oxo acid dehydrogenase acyltransferase, catalytic domain	3.00E^{-06}
pSB0079	A	AU100962	AC004684	Aldo/keto reductase family; dehydrogenase	5.00E^{-07}
pSB1362	B	AU091281	T02192	Cytochrome P450	8.00E^{-19}
		BE405309			3.00E^{-04}
pSB1059	C	AW745117	U56731	Phytochrome	4.00E^{-59}
pSB0615	F	BE639074	T02002	DHHC zinc finger domain	5.00E^{-03}
pSB1473	A	BE99566	AAD12714	PHD zinc finger	2.00E^{-03}
pSB1464	H	AW745146	D13513	Fructose-bisphosphate aldolase class-I	4.00E^{-35}
		BE597480			7.00E^{-47}
pSB1736	D	BE362296	S58286	Gamma-glutamyltranspeptidase	5.00E^{-03}
pSB1472	C	AU070926	T04207	Glutathione peroxidase	1.00E^{-04}
pSB0062	C	BE362533	AAF22517	Glutathione S-transferase (GST)	5.00E^{-08}
pSB1394	F	BE595814	AF050129	Glycosyl hydrolase family 32; cell wall invertase	4.00E^{-06}
pSB1379	D	AW678137	O04885	Glyoxalase	2.00E^{-10}
pSB0455	F	BE498242	T05741	Heat shock protein (HSP70)	2.00E^{-25}
pSB0720	B	AI881920	Y08987	Herpesvirus Glycoprotein B	4.00E^{-03}
pSB0487	D	AW057092	U04434	Iron/Ascorbate oxidoreductase family; senescence related	1.00E^{-13}
pSB1229	G	BE593740	X12540	Iron/manganese superoxide dismutase	7.00E^{-13}
pSB1654	E	BE357267	AAD22299	Homeobox domain	8.00E^{-03}
pSB0158	C	BE051782	AB003324	K-box region; floral homeotic protein; transcription factor	2.00E^{-10}
pSB1621	C	AI657258	T05943	Lipoxygenase	2.00E^{-37}
pSB0874	C	AA752663	AAC69138	Mitochondrial carrier protein	8.00E^{-09}
pSB1905	G	BE496970	AAC78333	Nitrogen regulatory protein P-II	4.00E^{-04}
pSB1460	D	BE598051	T04585	No apical meristem protein (NAM)	6.00E^{-16}
pSB0760	J	BE034304	T05448	Oxysterol-binding protein	3.00E^{-10}
pSB0613	A	BE238801	AB010074	Phosphoglucomutase/phosphomannomutase	5.00E^{-06}
pSB1108	J	BE444691	T05995	Polygalacturonase (pectinase)	6.00E^{-11}
pSB1633	I	AW057071	X66422	Polygalacturonase (pectinase)	2.00E^{-37}
pSB1233	A	BE592108	U08401	Prokaryotic-type carbonic anhydrase	8.00E^{-14}
pSB1163	J	BE360347	U92540	Proteasome A-type and B-type; endopeptidase beta chain	1.00E^{-19}
pSB0140	D	BE361635	AAF07388	PTR2, POT family; oligopeptide transport	4.00E^{-21}
pSB1600	H	AU075497	AAC28086	PTR2, POT family; oligopeptide transport	2.00E^{-13}
pSB0108	J	AW922460	CAB71132	Purple acid phosphatase	3.00E^{-07}
pSB1436	D	AW671157	F71418	B56 family protein phosphatase 2A regulatory B subunit	8.00E^{-34}
pSB0635	F	BE517956	CAB92051	Trehalose phosphatase	9.00E^{-06}
pSB1766	B	AI855362	CAB92051	Trehalose phosphatase	2.00E^{-14}
pSB0097	C	BE434674	AAF22889	Serine/Threonine protein phosphatase	2.00E^{-16}
pSB1777	C	AW424701	P52711	Serine carboxypeptidase	1.00E^{-18}
pSB1913	B	BE025336	AAB71481	Serine carboxypeptidase	5.00E^{-05}
pSB0164	J	AU057426	AAD25546	Eukaryotic protein kinase, catalytic domain	5.00E^{-06}
pSB0061	G	BE213652	AAF04884	Serine/threonine protein kinase, catalytic domain	3.00E^{-08}
pSB0077	B	BE366392	T04832	Serine/threonine protein kinase, catalytic domain	8.00E^{-14}
pSB0142	I	BE129698	T04109	Serine/threonine protein kinase, catalytic domain	2.00E^{-08}
pSB0182	E	AW433410	AAF68126	Serine/threonine protein kinase, catalytic domain	7.00E^{-04}
pSB0240	H	BE596011	T04832	Serine/threonine protein kinase, catalytic domain	5.00E^{-07}
pSB0289	A	AW042251	AAD21713	Serine/threonine protein kinase, catalytic domain	3.00E^{-12}
pSB0543	H	AW066532	T04109	Serine/threonine protein kinase, catalytic domain	1.00E^{-09}
pSB1140	A,C,G	AI649601	U95973	Serine/threonine protein kinase, catalytic domain	3.00E^{-09}
pSB0023	G	BE363874	AC005315	Tyrosine kinase, catalytic domain	5.00E^{-09}
pSB0435	F	AW923267	S61766	Tyrosine kinase, catalytic domain	3.00E^{-06}
pSB0783	B	AI944224	AAD39286	Tyrosine kinase, catalytic domain	1.00E^{-08}
pSB1236	I	AI491492	AAF27131	Tyrosine kinase, catalytic domain	3.00E^{-09}
pSB1254	J	AI881708	S71477	START domain; Star-related lipid transport domain	1.00E^{-05}
pSB0688	A	C25300	AP002483	SWI3, ADA2, N-CoR and TFIIIB DNA binding domains	1.00E^{-03}
pSB0860	H	BE040062	X78846	SWI3, ADA2, N-CoR and TFIIIB DNA binding domains	2.00E^{-05}
pSB1462	A	AW676687	AAF27693	Aminotransferase class-I	4.00E^{-06}
pSB1916	D	AW924285	T07131	Transmembrane amino-acid transporter protein	2.00E^{-18}
pSB1591	A	T75723	AC012193	Transmembrane amino-acid transporter protein; permease	9.00E^{-03}
pSB1433	A,B	AI820414	AJ277097	TruB family pseudouridylate synthase; Uracil hydrolase; kinetochore binding	2.00E^{-12}
pSB1814	C	BE595271	D29718	Ubiquitin-conjugating enzyme	1.00E^{-13}
pSB0896	F	AW400053	S52003	Major intrinsic protein (MIP)	1.00E^{-16}
pSB1310	D	BE366913	T02939	Voltage gated chloride channels	8.00E^{-31}

[a] Linkage group designations follow Chittenden et al. 1994
[b] E-value from CDD search

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 (D_{avg} = 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, previ-

Table 3 SSRs derived from RFLP probe sequences[a]

SSR	Probe name	LG[d]	Repeat	Forward primer (5′–3′)	Reverse primer (5′–3′)	No. alleles	D
Xcup01[b]	pSB0041	C	$(GA)_8$	*TET*-CATGGGCGGGTTGAAGAC	TGCAGGAAGGGAGGATGTAG	3	0.4305
Xcup02	pSB0069	G	$(GCA)_6$	*TET*-GACGCAGCTTTGCTCCTATC	GTCCAACCAACCCACGTATC	5	0.6925
Xcup05	pSB0094	F	$(GA)_8$	*HEX*-GGAAGGTTTGCAAGAACAGG	CCAGCCCAACAAGTGCTATC	7	0.7920
Xcup06	pSB0105	C	$(CTGC)_4$	*HEX*-GGCAGTAGCAGGCGTTTAAG	AACTGAATCAGGTCATGGGC	2	0.5227
Xcup07[bc]	pSB0115	I	$(CAA)_8$	*FAM*-CTAGAGGATTGCTGGAAGCG	CTGCTCTGCTTGTCGTTGAG	5	0.5699
Xcup08	pSB0130	Unmapped	$(TG)_6$	*TET*-GCAGCAACCACTTCCGATTC	GCAGTGCCGTCAAAAAGTAG	2[e]	0.2223
Xcup09	pSB0204	J	$(GAAT)_4$	*FAM*-CTGGTGAGGACAGCACAATG	CTTCTTGCCTATCTCTGCCC	1	0.0000
Xcup11[b]	pSB1889	A	$(GCTA)_4$	*TET*-TACCGCCATGTCATCATCAG	CGTATCGCAAGCTGTGTTTG	3	0.5681
Xcup12[b]	pSB1824	D	$(TG)_7$	*TET*-TGTTACAGAGACGCGCAGAG	GGCTGGTTGCTACCTTGTTC	4	0.5842
Xcup13[b]	pSB1810	I	$(CCGG)_5$	*HEX*-TCTCCTCCACCTTGTCAACC	CCTTGCCATCGACCACTC	2	0.4748
Xcup14	pSB1802	A	$(AG)_{10}$	*HEX*-TACATCACAGCAGGGACAGG	CTGGAAAGCCGAGCAGTATG	6	0.5475
Xcup15[c]	pSB1790	C	$(TCCCC)_4$	*FAM*-ATACACTCCCAAGCCAGCAC	CAATAAAAGAAGGGGGGAGC	3	0.2608
Xcup16[b]	pSB1771	I	$(CTTTT)_4$	*FAM*-TGCAGTGCTAGCTCATGGTC	CTTTCCAGCCTCCCATATCC	2	0.4200
Xcup17	pSB1764	D,J	$(AGC)_5$	*FAM*-CTGAGGAGTGGTTTCATCCC	CATCACCGTTCCCCTCTTC	2	0.0408
Xcup18[c]	pSB1757	G	$(CAAG)_4$	*FAM*-GCCTCTACATATCCACAAGCC	CAGATCAGTCATGCCACCTG	1	0.0000
Xcup19[bc]	pSB1755	J	$(CG)_7$	*FAM*-CCGAGTTCTCACTCCCTCTC	GACCTTGTCGAACTGCTTCC	2	0.0800
Xcup20	pSB1703	F	$(AT)_6$	*HEX*-TGGGTGTTGCACTGGGAG	ACTGAAAGCACCGTCTCTGG	3	0.4067
Xcup21	pSB1007	E	$(GAT)_5$	*FAM*-ATACCATCCACCTCACCAGC	GAAACGTACATGGGTTTGGG	1	0.0000
Xcup22[c]	pSB1016	C	$(AGTAC)_4$	*FAM*-CCAGTTCAGTTCAGTCCATACG	CGACAGCGCACACAAGTC	2	0.0801
Xcup23[b]	pSB1060	F	$(GCT)_5$	*TET*-GATAACTTTGGCCAACTCGC	TGTCTGCCCAGTTCCCAC	2	0.4800
Xcup24	pSB1126	C	$(TA)_9$	*HEX*-AAACTGGATGCCACACCAAG	AGCTATACCAACACGGGCAG	5[e]	0.8120
Xcup25	pSB1129	Unmapped	$(ACG)_5$	*TET*-GACACCGTGCAAAGGATAGC	GCACCAAAGCAGTTCCAGTG	2	0.3929
Xcup26[b]	pSB1144	B	$(CT)_6$	*HEX*-CGATCATCAGATCATGGGAG	CACTTGGGAAGTTGGGATTG	2	0.3575
Xcup27[c]	pSB1172	C	$(CT)_6$	*HEX*-AGAAGGACGACGAGAAGCAG	TGGAAGAGTACGGATCGAGG	2	0.0800
Xcup28[c]	pSB1217	F	$(TGAG)_5$	*TET*-GGTGTGAGACTGTGAGCAGC	TATAGCACGGTTGTTGTGCC	4	0.6042
Xcup29[c]	pSB1253	B	$(AT)_6$	*HEX*-CTTTCTCGATTTCTGGTGCC	TTTACCTTGCCTATGCCTGC	3	0.6225
Xcup32	pSB1359	A	$(AAAAT)_4$	*TET*-ACTACCACCAGGCACCACTC	GTACTTTTTCCCTGCCCTCC	3	0.4925
Xcup33[b]	pSB1431	C	$(AT)_7$	*FAM*-GCGCTGCTGTGTGTTGTTC	ACGGGGATTAGCCTTTTAGG	5	0.4492
Xcup34[c]	pSB1433	A,B	$(TTC)_5$	*TET*-GCCTCAGCTGACTCCAATTC	CTGATGTTTCTGTTCCTGCG	1	0.0000
Xcup36	pSB1456	B,D	$(CA)_8$	*TET*-TGAGCTGATAATGGCTGCTG	GCGTCACGGAAGTTGGAC	5	0.4700
Xcup37[bc]	pSB1460	D	$(AG)_9$	*HEX*-CCCAGCCTTCCTCCTGATAC	GTACCGACTCCAATCCAACG	3	0.2267
Xcup38	pSB1463	C	$(ACT)_5$	*TET*-CTCTCACGGAAAGGAAGCAC	TACCGAAGCGGAAGCTACTC	2	0.2800
Xcup40	pSB1490	B	$(TG)_7$	*HEX*-ACGGAGAATAGAAAGTGGCG	TTGAGCATGCAACCACCTAC	3	0.6108
Xcup41	pSB1497	B	$(CAA)_5$	*FAM*-AACACGAAAAGGTTAGGGGG	TCGAATGGTCCAGTAGTCCC	1	0.0000
Xcup42[c]	pSB1499	I	$(GA)_9$	*TET*-CACACCTGTCTCTCTTCTCCG	AGATCATCTTCGCCTTCCTC	2	0.2200
Xcup43[b]	pSB1511	I	$(CTGCC)_5$	*FAM*-GCCTAACTCCCTTGTGATGC	GTCAGTGGATGTGGATGTGC	2	0.4675
Xcup44	pSB1520	Unmapped	$(AC)_6$	*HEX*-CATGCATGCGTGTACCTGAG	TAGCTGTGTCCGTCGATGTC	1	0.0000
Xcup47[c]	pSB1549	E	$(GA)_{21}$	*FAM*-TGAGCAATGAACTTAGGGGG	CTACCCTTTGATGGCAGTACC	6	0.7508
Xcup48[bc]	pSB1565	F	$(AT)_7$	*FAM*-TCACTAGCGCCTCCAAAATC	TCCAATCCTTCCTGTGCTTC	7	0.6842
Xcup49[b]	pSB0305	I	$(GGAT)_6$	*TET*-TCCACCTCCATCATCTTTCC	CTCCACCACCTCCATGACTC	6	0.6233
Xcup50[b]	pSB0305	I	$(ACAGG)_5$	*TET*-TGATTGATTGAGGCAGGCAC	TTCCGGTCTCTGTCCATTTC	4	0.5767
Xcup52[b]	pSB0491	J	$(AATT)_5$	*FAM*-CTCCTCGCCGTCATCATC	TAAAGAGAAACGCAGGCAGG	3	0.5300
Xcup53[b]	pSB0508	C	$(TTTA)_5$	*HEX*-GCAGGAGTATAGGCAGAGGC	CGACATGACAAGCTCAAACG	4	0.6075
Xcup55	pSB0528	A	$(CGC)_5$	*FAM*-AGCTGCTCTGCTTCCAGTTC	TCTTCGTCAACGTGCTCATC	1	0.0000
Xcup57[bc]	pSB0540	J	$(TAGC)_5$	*TET*-CTGCAGAGAGCTAATTGTGC	TCTTGGAAGAGACGGACCTG	4	0.4986
Xcup58	pSB0054	B	$(GATC)_4$	*TET*-TAGAGCTGATCGAGGGATGG	AGCTAGCCGACACCAACATC	1	0.0000
Xcup60	pSB0558	C	$(CGGT)_4$	*TET*-GTATGCATGGATGCCTGATG	GCGAGGGTATGTAGCTCGAC	2	0.0800
Xcup61[c]	pSB0581	A	$(CAG)_7$	*HEX*-TTAGCATGTCCACCACAACC	AAAGCAACTCGTCTGATCCC	2	0.5175
Xcup62[c]	pSB0600	C	$(GAA)_6$	*HEX*-CGAGAAGATCGAGAGAACCC	TGAAGACGACGACGACAGAC	2	0.4375
Xcup63[b]	pSB0605	B	$(GGATGC)_4$	*TET*-GTAAAGGGCAAGGCAACAAG	GCCCTACAAAATCTGCAAGC	2	0.1533
Xcup64[c]	pSB0606	B	$(TA)_9$	*HEX*-TATTGACACGCAGGTAACGC	GAGGACGAGTGCATGATGAG	4	0.6092
Xcup65[c]	pSB0613	A	$(AAAC)_4$	*HEX*-GCAATTGACAACGCATCTGG	AGTAATCGTCTCCGGTGCTG	1	0.0000
Xcup66	pSB0632	I	$(AT)_6$	*FAM*-GGCTTTAGCGATCGAGCTTC	AGGGTACGACGTGGAGATTG	2[e]	0.3889
Xcup67[bc]	pSB0703	I	$(TA)_6$	*FAM*-GGTCAGTGCTTACACAGATTCC	GGGGATTGCAGGTGTCATAG	4	0.7192
Xcup68[bc]	pSB0716	J	$(TGAT)_5$	*HEX*-TACCTCACCCACTCCTACCG	AACCTCACCTGCAATCAACC	2	0.2800
Xcup69[bc]	pSB0720	B	$(ATGCG)_4$	*FAM*-ACAGCACCAAGGTGAAGGAC	ATGTAGGGCACCAGCTTCAC	3	0.4925
Xcup70[b]	pSB0815	J	$(TTGTT)_5$	*TET*-GGAGGAACACGCACAAAAAG	CACTCTAGCTATGGCCTGGG	4	0.3567
Xcup71[bc]	pSB0896	F	$(CA)_7$	*TET*-CCACCTGTTGATGGGTTCC	AGCTTCGTCGTCTCTGGTTC	2	0.3800
Xcup73[c]	pSB0948	C	$(TA)_{10}$	*HEX*-GGTTCTGTCGTCATCACCAG	ATCTTTAGCCGCCACATGAC	6	0.8525
Xcup74	pSB0986	B	$(TG)_9$	*FAM*-GTCGCCATTGTGATGAAGAG	CAGTAGTCCAGCAAAACGGC	5	0.5133

[a] Allele size range, number of alleles, and D are for lines 1–25 only (Table 1

[b] Allele sizes do not conform to stepwise mutation model

[c] BLAST E-value significant

[d] Linkage group destinations follow Chittenden et al. 1994

[e] > 10% null alleles

Table 4 Information content of SSR loci assayed in inbreds and sorghum diverse lines

SSR	Number of alleles			Diversity index (D)		
	Inbreds[a]	Diverse[b]	All lines	Inbreds	Diverse	All lines
Xcup01	3	10	10	0.4274	0.7807	0.7002
Xcup02	5	4	5	0.6631	0.6638	0.6847
Xcup05	6	11	14	0.7869	0.8165	0.8408
*Xcup06**	2	3	3	0.5250	0.4041	0.4707
Xcup07	5	8	9	0.5524	0.7909	0.7571
Xcup12	4	7	8	0.6083	0.7617	0.7217
*Xcup13**	2	3	3	0.5060	0.2542	0.3658
Xcup14	4	7	8	0.4440	0.6170	0.5733
Xcup16	2	3	3	0.3238	0.5585	0.5054
Xcup23	2	2	2	0.4667	0.4002	0.6239
Xcup25	2	3	3	0.2412	0.3316	0.3031
Xcup28	4	4	4	0.6190	0.6265	0.6223
Xcup29	3	3	3	0.6417	0.5611	0.5917
*Xcup32**	3	2	3	0.5131	0.2837	0.3597
*Xcup33**	5	4	5	0.5179	0.3810	0.4155
Xcup36	4	5	5	0.4714	0.6411	0.5937
Xcup40	3	4	4	0.6321	0.7342	0.7092
Xcup43	2	4	4	0.4821	0.6573	0.6130
*Xcup47**	5	5	6	0.7583	0.5506	0.6406
Xcup49	6	8	9	0.6917	0.7963	0.8115
Xcup50	3	6	6	0.5524	0.7692	0.7727
Xcup52	2	4	4	0.4952	0.5122	0.5423
Xcup53	4	5	6	0.6345	0.6373	0.6761
Xcup57	4	6	6	0.4561	0.6769	0.6224
Xcup61	2	2	2	0.5202	0.5009	0.5045
Xcup66	2	2	2	0.4762	0.4971	0.4850
Xcup69	3	5	5	0.4750	0.6491	0.6403
Xcup73	6	8	8	0.8369	0.8150	0.8133
Xcup74	5	4	5	0.4714	0.6752	0.6376
Average	3.6	4.9	5.3	0.5445	0.5981	0.6068

[a] Table 1, nos. 1–22
[b] Table 1, nos. 23–37
* D (inbred lines) > 10% D (diverse lines)

ously 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, $D_{avg} = 0.46$).

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 de-

Table 5 SSRs located in or near putative genes

SSR	EST/nt, nr[a] accession	Putative gene identification	Function	SSR position
Xcup07	None/AC006434	*Arabidopsis thaliana* tRNA threonine	tRNA	Upstream
Xcup15	C98839/None	*Oryza sativa* EST	Unknown	Unknown
Xcup18	AU076044/CAA18115	*A. thaliana* hypothetical protein	Unknown	Downstream of stop
Xcup19	None/AAD21445	*A. thaliana* putative kinesin-related cytokinesis protein	Microtubule motor	Upstream of start
Xcup20	AW679270/None	*S. bicolor* EST	Unknown	Unknown
Xcup22	AI665315/None	*Z. mays* EST	Unknown	Unknown
Xcup27	None/AF061282	*S. bicolor* 22 kDa kafirin cluster	Seed storage proteins and TEs	Unknown
Xcup28	AW065891/CAB88990	*A. thaliana* hypothetical protein	Unknown	Upstream of start
Xcup29	AW285130/None	*S. bicolor* EST	Unknown	3′ UTR
Xcup34	AI820414/AJ277097	*Z. mays* kinetochore binding protein	Chromosome movement	Coding region
Xcup37	None/AAD20120	*A. thaliana* NAM-like protein	Developmental regulator	Intron
Xcup42	AW681032/AP001168	*O. sativa* hypothetical protein	Unknown	Coding region
Xcup47	BE344922/AAD55604	*A. thaliana* signal peptidase	Organellar localization of proteins	3′ UTR
Xcup48	None/AAD21414	*A. thaliana* terpene synthesis protein	Terpene synthesis	Intron
Xcup57	None/CAB89322	*A. thaliana* hypothetical protein	Unknown	Intron
Xcup61	BE475831/P55195	*Vigna aconitifolia* phosphoribosyl-aminoimidazole carboxylase	Purine biosynthesis	Intron
Xcup62	AW566364/None	*Z. mays* EST	Unknown	Unknown
Xcup64	AW746789/None	*A. thaliana* SH27A-like protein	Transporter	3′ UTR
Xcup65	BE238801/CAC09322	*Pisum sativum* plastidic phosphoglucomutase	Starch synthesis	Coding region
Xcup67	BE051702/AAC99309	*Malus domestica* Constans-like protein	Developmental regulator	Intron
Xcup68	D24867/AAF79837	*A. thaliana* carbonic anhydrase	Oxygen exchange in chloroplast	Intron
Xcup69	AI881920/Y08987	*O. sativa* osr40g2 gene	Osmotic stress response	Intron
Xcup71	AW400053/S52003	*O. sativa* major intrinsic protein	Ion channel	5′ UTR
Xcup73	AW744875/T01059	*A. thaliana* lupeol synthase	Sterol synthesis	5′ UTR

[a] nt – nucleotide database, nr – protein database

tected 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 hexanucleotide repeats, and these markers were almost as informative as the di-nucleotide SSRs (D_{avg} = 0.43 compared to D_{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 ≥ 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.

References

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Anderson I, Brass A (1998) Searching DNA databases for similarities to DNA sequences: when is a match significant? Bioinformatics 14:349–356

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Pl Mol Biol Rep 9:208–218

Bhattramakki D, Dong JM, Chhabra AK, Hart GE (2000) An integrated SSR and RFLP linkage map of *Sorghum bicolor* (L.) Moench. Genome 43:988–1002

Bowers JE, Schertz KF, Abbey C, Anderson S, Chang C, Chittenden LM, Draye X, Hoppe AH, Jessup R, Lennington J, Paterson AH (2000) A high-density 2,399-locus genetic map of *Sorghum*. Plant and Animal Genome VIII Conference, San Diego, (http://www.intl-pag.org/pag/8/abstracts/pag8712.html)

Brown SM, Hopkins MS, Mitchell SE, Senior ML, Wang TY, Duncan RR, Gonzalez-Candelas F, Kresovich S (1996) Multiple methods for the identification of polymorphic simple sequence repeats (SSRs) in sorghum [*Sorghum bicolor* (L.) (Moench)]. Theor Appl Genet 93:190–198

Chen MS, SanMiguel P, Bennetzen JL (1998) Sequence organization and conservation in *sh2/a1*-homologous regions of sorghum and rice. Genetics 148:435–443

Chittenden LM, Schertz KF, Lin YR, Wing RA, Paterson AH (1994) A detailed RFLP map of *Sorghum bicolor* ×*S. propinquum*, suitable for high-density mapping, suggests ancestral duplication of sorghum chromosomes or chromosomal segments. Theor Appl Genet 87:925–933

Condit R, Hubbell SP (1991) Abundance and DNA sequence of two-base repeat regions in tropical tree genomes. Genome 34:66–71

Daly MJ, Lincoln SE, Lander ES (1991) "PRIMER", unpublished software, Whitehead Institute/MIT Center for Genome Research

Davis GL, McMullen MD, Baysdorfer C, Musket T, Grant D, Staebell M, Xu G, Polacco M, Koster L, Melia-Hancock S, Houchins K, Chao S, Coe EH (1999) A maize map standard with sequenced core markers, grass genome reference points and 932 expressed sequence tagged sites (ESTs) in a 1,736-locus map. Genetics 152:1137–1172

Doggett H (1988) Sorghum, 2nd edn. John Wiley and Sons Inc, New York

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small amounts of leaf tissue. Phytochem Bull 19:11–15

Gaut BS, Doebley JF (1997) DNA sequence evidence for the segmental allotetraploid origin of maize. Proc Natl Acad Sci USA 94:6809–6814

Harlan JR, de Wet JMJ (1972) A simplified classification of cultivated sorghum. Crop Sci 12:172–176

Harushima Y, Yano M, Shomura P, Sato M, Shimano T, Kuboki Y, Yamamoto T, Lin SY, Antonio BA, Parco A, Kajiya H, Huang N, Yamamoto K, Nagamura Y, Kurata N, Khush GS, Sasaki T (1998) A high-density rice genetic linkage map with 2,275 markers using a single F-2 population. Genetics 148:479–494

Innan H, Terauchi R, Miyashita T (1997) Microsatellite polymorphism in natural populations of the wild plant *Arabidopsis thaliana*. Genetics 146:1441–1452

Kimura M, Crow JF (1964) The numbers of alleles that can be maintained in a finite population. Genetics 49:725–738

Kong L, Dong J, Hart GE (2000) Characteristics, linkage-map positions, and allelic differentiation of *Sorghum bicolor* (L.) Moench DNA simple-sequence repeats (SSRs). Theor Appl Genet 101:438–448

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4:203–221

Matsuoka Y, Mitchell SE, Kresovich S, Goodman M, Doebley J (2002) Microsatellites in *Zea*-variability, patterns of mutations, and use for evolutionary studies. Theor Appl Genet 104:436–450

Michalek W, Kunzel G, Graner A (1999) Sequence analysis and gene identification in a set of mapped RFLP markers in barley (*Hordeum vulgare*). Genome 42:849–853

Ming R, Liu SC, Lin YR, da Silva J, Wilson W, Braga D, van Deynze A, Wenslaff TF, Wu KK, Moore PH, Burnquist W, Sorrells ME, Irvine JE, Paterson AH (1998) Detailed alignment of *Saccharum* and *Sorghum* chromosomes: comparative organization of closely related diploid and polyploid genomes. Genetics 150:1663–1682

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70:3321–3323

Peng Y, Schertz DF, Cartinhour S, Hart GE (1999) Comparative genome mapping of *Sorghum bicolor* (L.) Moench using an RFLP map constructed in a population of recombinant inbred lines. Plant Breed 118:225–235

Schug MD, Hutter CM, Wetterstrand KA, Gaudette MS, Mackay TFC, Aquadro CF (1998) The mutation rates of di-, tri-, and tetra-nucleotide repeats in *Drosophila melanogaster*. Mol Biol Evol 5:1751–1760

Senior ML, Murphy JP, Goodman MM, Stuber CW (1998) Utility of SSRs for determining genetic similarities and relationships in maize using an agarose-gel system. Crop Sci 38:1088–1098

Ventelon M, Deu M, Garsmeur O, Doligez A, Ghesquiere A, Lorieux M, Rami JF, Glaszmann JC, Grivet L (2001) A direct comparison between the genetic maps of sorghum and rice. Theor Appl Genet 102:379–386

Weber JL (1990) Informativeness of human (dC-dA)*n* (dG-dT)*n* polymorphisms. Genomics 7:524–530