Annals of Botany **105**: 471–480, 2010
doi:10.1093/aob/mcp305, available online at www.aob.oxfordjournals.org

ANNALS OF BOTANY
Founded 1887

Phylogenetic analysis of the genus *Sorghum* based on combined sequence data from cpDNA regions and ITS generate well-supported trees with two major lineages

Dickson Ng'uni[1,*], Mulatu Geleta[1], Moneim Fatih[2] and Tomas Bryngelsson[1]

[1]*Department of Plant Breeding and Biotechnology, Swedish University of Agricultural Sciences, Box 101, SE 230 53 Alnarp, Sweden and* [2]*Swedish Biodiversity Centre, Box 91, Pomonavågen 2, SE-230 53 Uppsala, Sweden*
** For correspondence. E-mail dickson.nguni@gmail.com*

Received: 3 October 2009 Returned for revision: 3 November 2009 Accepted: 26 November 2009 Published electronically: 7 January 2010

• *Background and Aims* Wild *Sorghum* species provide novel traits for both biotic and abiotic stress resistance and yield for the improvement of cultivated sorghum. A better understanding of the phylogeny in the genus *Sorghum* will enhance use of the valuable agronomic traits found in wild sorghum.
• *Methods* Four regions of chloroplast DNA (cpDNA; *psb*Z-*trn*G, *trn*Y-*trn*D, *trn*Y-psbM and *trn*T-*trn*L) and the internal transcribed spacer (ITS) of nuclear ribosomal DNA were used to analyse the phylogeny of sorghum based on maximum-parsimony analyses.
• *Key Results* Parsimony analyses of the ITS and cpDNA regions as separate or combined sequence datasets formed trees with strong bootstrap support with two lineages: the *Eu-sorghum* species *S. laxiflorum* and *S. macrospermum* in one and *Stiposorghum* and *Para-sorghum* in the other. Within *Eu-sorghum*, *S. bicolor*-3, -11 and -14 originating from southern Africa form a distinct clade. *S. bicolor*-2, originally from Yemen, is distantly related to other *S. bicolor* accessions.
• *Conclusions Eu-sorghum* species are more closely related to *S. macrospermum* and *S. laxiflorum* than to any other Australian wild *Sorghum* species. *S. macrospermum* and *S. laxiflorum* are so closely related that it is inappropriate to classify them in separate sections. *S. almum* is closely associated with *S. bicolor*, suggesting that the latter is the maternal parent of the former given that cpDNA is maternally inherited in angiosperms. *S. bicolor*-3, -11 and -14, from southern Africa, are closely related to each other but distantly related to *S. bicolor*-2.

Key words: Molecular phylogeny, *Sorghum*, *Eu-sorghum*, *Zea mays*, non-coding regions, cpDNA, ITS.

INTRODUCTION

Sorghum Moench is highly heterogeneous and with *Cleistachne* Bentham form Sorghastrae (Garber, 1950), one of the 16 subtribes belonging to tribe Andropogoneae. Species of the genus *Sorghum* have chromosome numbers of $2n = 10$, 20, 30 or 40 (Garber, 1950; Lazarides *et al.*, 1991). There are five recognized sections and 25 species within *Sorghum.* The sections are *Eu-sorghum*, *Chaetosorghum*, *Heterosorghum*, *Para-sorghum* and *Stiposorghum* (Garber, 1950; Lazarides *et al.*, 1991). *Eu-sorghum* includes cultivated sorghums and their closest wild relatives (De Wet and Huckay, 1967). According to De Wet (1978) three species were recognized in section *Eu-sorghum*, including two perennial species, *S. halepense* and *S. propinquum*, and an annual, *S. bicolor*. However, in the earlier classification by Snowden (1935), *Eu-sorghum* is considered to comprise two subsections, *Arundinacea* and *Halepensia.* The subsection *Arundinacea*, commonly found in tropical Africa and India, consists of *S. bicolor* (L.) Moench, *S. arundinaceum* (Desv.) Stapf and *S. drummondii* (Steud.) Millsp. *S. propinquum* (Kunth) Hitchcock, *S. halepense* (L.) Pers and *S. almum* Parodi form subsection *Halepensia*, and are found in the Mediterranean region and Southeast Asia.

The wild Australian *Sorghum* species constitute over two-thirds of the recognized *Sorghum* species, of which one species each belongs to *Chaetosorghum* and *Heterosorghum.* The section *Para-sorghum* comprises seven species. Of these, five are native to northern monsoonal Australia, Africa and Asia (Garber, 1950; Lazarides *et al.*, 1991). *Stiposorghum* consists of ten species that are endemic to northern Australia (Garber, 1950; Lazarides *et al.*, 1991). The wild and weedy *Sorghum* species present a valuable source of agronomic traits such as pest and disease resistance (Sharma and Franzmann, 2001; Kamala *et al.*, 2002; Komolong *et al.*, 2002) for introgression into *S. bicolor*. Exploitation of these valuable traits requires a thorough understanding of the phylogenetic relationships between cultivated sorghum and the wild sorghum gene pool.

The chloroplast genome is useful in providing information on the inference of the evolutionary patterns and processes in plants (Raubeson and Jansen, 2005). The genome has, either solely or combined with other genomes, been widely used for inferring phylogenetic relationships of different taxa, including *Hordeum*, *Triticum* and *Aegilops* (Gielly and Taberlet, 1994), *Guizotia* (Geleta, 2007), Solanaceae (Melotto-Passarin *et al.*, 2008) and *Sorghum* (Dillon *et al.*, 2007). The non-coding chloroplast regions are phylogenetically more informative than the coding regions at lower taxonomic levels because they are under less functional constraints and evolve rapidly (Gielly and Taberlet, 1994). One of the

chloroplast DNA (cpDNA) regions, *trn*T-*trn*L, used in this study was reported to possess sufficient phylogenetic signal for studies at lower taxonomic levels (Shaw *et al.*, 2005).

The internal transcribed spacer (ITS) region of the 18S–5·8S–26S nuclear ribosomal DNA (nrDNA) has been commonly used for phylogenetic inference at the generic and infrageneric level in plants. The ITS loci properties of biparental inheritance, universality of primers, intragenomic uniformity and intergenomic variability merit their utility for phylogenetic reconstruction (Baldwin *et al.*, 1995). Two ITS regions, ITS1 and ITS2, generally evolve more rapidly than coding regions and have shown to be equally informative, being able to differentiate between closely related species (Baldwin, 1992) and more specifically to resolve phylogenetic relationships of sorghum and related species (Sun *et al.*, 1994; Dillon *et al.*, 2001; Guo *et al.*, 2006).

This study sought to resolve the phylogenetic relationships between species of the genus *Sorghum* based on four regions of the cpDNA: *trn*Y-*trn*D, *psb*Z- *trn*G, *trn*Y-*psb*M and *trn*T-*trn*L and the ITS of nrDNA and also to evaluate the usefulness of the five non-coding regions of cpDNA in resolving relationships among the closely related species within section *Eu-sorghum*.

MATERIALS AND METHODS

Plant material

Details of twenty-two *Sorghum* species along with GenBank germplasm and GenBank sequence accession numbers used in this study are given in Table 1. The germplasm accessions included wild sorghum and several cultivated sorghum obtained from the Australian Tropical Crops Genetic Resource Centre, Biloela, Queensland, Australia. In addition, five accessions of *S. bicolor* and one of *S. arundinaceum* were obtained from the Zambian National Plant Genetic Resources Centre.

DNA extraction, PCR and sequencing

Each *Sorghum* species was represented by 1–2 accessions, except for *S. bicolor* for which 11 accessions were used. Genomic DNA was extracted from fresh leaf tissues of seedlings raised in the greenhouse at approx. 2 weeks of age using a modified CTAB extraction method (Doyle and Doyle, 1987). The quality of the DNA was analysed by agarose gel electrophoresis and DNA concentration was determined using a Nanodrop® ND-1000 spectrophotometer (Saveen Werner, Malmö, Sweden).

Primers for amplification and sequencing of the *trn*S-*trnf*M, *trn*Y-*psb*M and *trn*T-*trn*D regions were designed for this study while the *trn*T-*trn*L region was amplified and sequenced using the universal primers designed by Taberlet *et al.* (1991). A primer pair was used for each of the cpDNA regions. However, two primer pairs were designed for amplification of the *trn*Y-*psb*M region. Universal primers ITS4 and ITS5 (White *et al.*, 1990) were used for amplification and sequencing of the ITS region.

The sequences of the primers and information on specific primers supplied by Eurofins MWG GmbH (Ebersberg, Germany) used in this study are given in Table 2. A GeneAMP PCR system 9700 thermocycler was used for amplification with the following temperature regime: denaturation at 94 °C, followed by 30 cycles of 1 min denaturing at 94 °C, 1 min primer annealing at 51 °C and primer extension for 2 min at 72 °C, and a final 7-min extension at 72 °C. Successfully amplified samples were purified using the QIAquick PCR purification kit (Qiagen GmbH, Hilden, Germany) and microcentrifuge according to the manufacturer's instructions. Nine microlitres of purified PCR products was mixed with 1 μL of sequencing primers and sent to the sequencing facility in the University of Oslo, Norway (http://www.bio.uio.no/ABI-lab/), where DNA sequencing was done. The quality of the sequences was evaluated using Sequence Scanner version 1.0 (Applied Biosystems, www.appliedbiosystems.com/) and only high-quality sequences were used for the analysis. All regions were sequenced using both forward and reverse primers. The sequences from the forward and reverse primers were aligned for each sample in order to generate a consensus sequence. As the sequences were of high quality, the forward and reverse sequences are identical, except in a few cases. These few discrepancies were resolved by repeated PCR and sequencing.

Sequence alignment and data analyses

The quality of the sequences was visually inspected using Sequence Scanner version 1.0 (Applied Biosystems). Multiple sequence alignment was performed using ClustalX version 2.1.10 (Larkin *et al.*, 2007). The sequences were edited using BioEdit version 7.0.9 (Hall, 1999) and PAUP* 4·0 Beta 10 was used for phylogenetic analyses. The phylogenetic analyses were approached in three ways. In the first approach, the ITS sequences of the nrDNA were analysed separately. In the second approach, the sequences of the four non-coding regions of the cpDNA were also analysed separately. In the final approach, a combined analysis of the cpDNA regions and the ITS was carried out. In all the cases indel positions were treated as missing data. *Zea mays* (GenBank accession no. U04796) was used as an out-group species.

RESULTS

Sequence characteristics of the Sorghum *species*

The sequence characteristics and parsimony-based tree statistics of four non-coding regions of cpDNA and the ITS are summarized in Table 3. The aligned sequences derived from all the cpDNA regions and the ITS revealed differences in sequence length between the *Sorghum* species. The longest sequences were obtained from the *trn*Y-*psb*M spacer and ranged from 1028 nt (*S. drummondii*) to 1053 nt (*S. exstans*). The eight *S. bicolor* sequences from this spacer exhibited 2–3 nt differences between them. By contrast, the *psb*Z-*trn*G spacer provided the shortest sequences, which ranged between 286 nt (*Eu-sorghum* species) and 291 nt (*S. intrans*). The similarity in sequence length between the *Eu-sorghum* species could be attributed to the occurrence of 5-nt indels within the *psb*Z-*trn*G intergenic spacer. Indels of similar size at corresponding positions were also observed in *S. laxiflorum* and *S. macrospermum.*

TABLE 1. *Accession identity and geographical origin of each accession of* Sorghum *species used in the study*

Species	Section	Germplasm accession no.*	DNA sequence accession no.				
			*trn*Y-*trn*D	*psb*Z-*trn*G	*trn*Y-*psb*M	*trn*T-*trn*L	ITS
S. almum	*Eu-sorghum*	AusTRCF302386^{A}	GQ121828	GQ121769	GQ121810	GQ121791	GQ121750
S. amplum-1	*Stiposorghum*	AusTRCF302455^{A}	N/A	N/A	N/A	N/A	N/A
S. amplum-2	*Stiposorghum*	AusTRCF302623^{A}	GQ121822	GQ121755	GQ121799	GQ121783	GQ121727
S. angustum-1	*Stiposorghum*	AusTRCF302588^{A}	GQ121824	N/A	GQ121793	GQ121775	GQ121737
S. angustum-2	*Stiposorghum*	AusTRCF302606^{A}	N/A	GQ121761	N/A	N/A	N/A
S. arundinaceum	*Eu-sorghum*	ZMB 7203^{Zm}	GQ121832	GQ121766	GQ121806	GQ121790	GQ121746
S. bicolor-1	*Eu-sorghum*	AusTRCF304111^{TA}	N/A	N/A	N/A	N/A	N/A
S. bicolor-2	*Eu-sorghum*	AusTRCF304113^{YA}	N/A	N/A	N/A	N/A	GQ121748
S. bicolor-3	*Eu-sorghum*	AusTRCF304114^{ZwA}	N/A	N/A	N/A	N/A	N/A
S. bicolor-4	*Eu-sorghum*	AusTRCF304115^{BA}	N/A	N/A	N/A	N/A	GQ121745
S. bicolor-5	*Eu-sorghum*	AusTRCF312813^{ZmA}	N/A	N/A	N/A	N/A	N/A
S. bicolor-14	*Eu-sorghum*	ZMB 5395^{Zm}	N/A	N/A	N/A	N/A	N/A
S. bicolor-12	*Eu-sorghum*	ZMB 5757^{Zm}	GQ121829	GQ121770	GQ121813	GQ121792	GQ121743
S. bicolor-15	*Eu-sorghum*	ZMB 6665^{Zm}	N/A	N/A	N/A	N/A	N/A
S. bicolor-10	*Eu-sorghum*	ZMB 7016^{Zm}	N/A	N/A	N/A	N/A	GQ121744
S. bicolor-11	*Eu-sorghum*	ZMB 7034^{Zm}	N/A	N/A	N/A	N/A	N/A
S. bicolor-13	*Eu-sorghum*	ZMB 7112^{Zm}	N/A	N/A	N/A	N/A	N/A
S. brachypodum-1	*Stiposorghum*	AusTRCF302480^{A}	GQ121818	GQ121756	GQ121802	GQ121774	GQ121736
S. brachypodum-2	*Stiposorghum*	AusTRCF302481^{A}	N/A	N/A	N/A	N/A	N/A
S. bulbosum-1	*Stiposorghum*	AusTRCF302418^{A}	N/A	N/A	N/A	N/A	N/A
S. bulbosum-2	*Stiposorghum*	AusTRCF302646^{A}	GQ121823	GQ121758	QG121803	GQ121781	GQ121732
S. drummondii-1	*Eu-sorghum*	AusTRCF300263^{EA}	N/A	N/A	N/A	N/A	N/A
S. drummondii-2	*Eu-sorghum*	AusTRCF300264^{KA}	GQ121831	GQ121765	GQ121809	GQ121789	GQ121747
S. ecarinatum-1	*Stiposorghum*	AusTRCF302450^{A}	GQ121821	GQ121754	GQ121800	GQ121784	GQ121730
S. ecarinatum-2	*Stiposorghum*	AusTRCF302662^{A}	N/A	N/A	N/A	N/A	N/A
S. exstans-1	*Stiposorghum*	AusTRCF302401^{A}	N/A	N/A	N/A	N/A	N/A
S. exstans-2	*Stiposorghum*	AusTRCF302473^{A}	GQ121816	GQ121759	GQ121796	GQ121782	GQ121735
S. halepense-1	*Eu-sorghum*	AusTRCF300167^{A}	GQ121830	GQ121768	GQ121808	GQ121788	N/A
S. halepense-2	*Eu-sorghum*	AusTRCF300188^{A}	N/A	N/A	N/A	N/A	GQ121749
S. interjectum-1	*Stiposorghum*	AusTRCF302396^{A}	GQ121817	GQ121753	GQ121797	GQ121772	GQ121738
S. interjectum-2	*Stiposorghum*	AusTRCF302433^{A}	N/A	N/A	N/A	N/A	N/A
S. intrans	*Stiposorghum*	AusTRCF302390^{A}	GQ121825	GQ121752	GQ121795	GQ121780	GQ121733
S. laxiflorum-1	*Heterosorghum*	AusTRCF302503^{A}	GQ121833	GQ121771	GQ1218011	GQ121786	GQ121741
S. laxiflorum-2	*Heterosorghum*	AusTRCF302607^{A}	N/A	N/A	N/A	N/A	N/A
S. leiocladum-1	*Para-sorghum*	AusTRCF300148^{A}	GQ121814	N/A	GQ121805	N/A	N/A
S. leiocladum-2	*Para-sorghum*	AusTRCF300170^{A}	N/A	GQ121763	N/A	GQ121778	GQ121739
S. macrospermum	*Chaetosorghum*	AusTRCF302367^{A}	GQ121834	GQ121767	GQ121812	GQ121787	GQ121742
S. matarankense-1	*Para-sorghum*	AusTRCF302521^{A}	GQ121826	GQ121757	GQ121804	GQ121776	GQ121731
S. matarankense-2	*Para-sorghum*	AusTRCF302636^{A}	N/A	N/A	N/A	N/A	N/A
S. nitidum-1	*Para-sorghum*	AusTRCF302539^{A}	N/A	N/A	N/A	GQ121785	N/A
S. nitidum-2	*Para-sorghum*	AusTRCF302558^{A}	GQ121815	GQ121764	GQ121807	N/A	GQ121740
S. plumosum-1	*Stiposorghum*	AusTRCF302399^{A}	GQ121819	GQ121762	GQ121798	N/A	N/A
S. plumosum-2	*Stiposorghum*	AusTRCF302489^{A}	N/A	N/A	N/A	GQ121773	GQ121729
S. plumosum-3	*Stiposorghum*	AusTRCF302635^{A}	N/A	N/A	N/A	N/A	N/A
S. stipoideum-1	*Stiposorghum*	AusTRCF302393^{A}	GQ121827	GQ121751	GQ121794	N/A	GQ121734
S. stipoideum-2	*Stiposorghum*	AusTRCF302669^{A}	N/A	N/A	N/A	GQ121779	N/A
S. timorense-1	*Para-sorghum*	AusTRCF302381^{A}	GQ121820	GQ121760	GQ121801	GQ121777	GQ121727
S. timorense-2	*Para-sorghum*	AusTRCF302459^{A}	N/A	N/A	N/A	N/A	N/A

N/A, not applicable.

* Superscripts at the end of the accession number denote the country of origin and the donor of that particular accession; if only a single country code is present then that country is both a donor and the origin of the accession. A, Australia; B, Burundi; E, Ethiopia; K, Kenya; T, Tanzania; Y, Yemen; Zm, Zambia; Zw, Zimbabwe.

Sequence length variations were also observed between *Sorghum* species in the *trn*T-*trn*L spacer, ranging from 684 nt (*S. arundinaceum*) to 693 nt (*S. leiocladum* and *S. laxiflorum*). Low sequence length differences of 2 nt in the *trn*T-*trn*L spacer were observed among the *S. bicolor* accessions. Significant sequence variations arising from transitions and transversions were observed at eight positions, which resulted in the discrimination of *S. bicolor*-12, -13 and -14 from the rest of the *S. bicolor* accessions. The sequences obtained from the *trn*Y-*trn*D spacer were between 318 nt (*S. amplum*, *S. angustum*) and 329 nt (*S. exstans*). The sequences obtained from the ITS showed narrow length differences of 528–534 nt between the *Sorghum* species; base substitutions in the ITS1 accounted for most of this variation. The *S. bicolor* accessions exhibited sequence length differences arising from a single nucleotide indel in the ITS1 region.

TABLE 2. *Primers used to amplify and sequence the five non-coding regions of cpDNA and the ITS of nrDNA*

Region of cpDNA	Primer name	Primer sequence (5′ → 3′)	Source of primer sequences
*psb*Z-*trn*G	tnSM –fw†	TGCTTCTCCTGATGGTTGGT	This study
	tnSM – rv†	GCTCGCTACATTGAACTACGC	
*trn*Y-*psb*M	psBD – fw*	CTGTCAAGGCGGAAGCTG	This study
	psBD – rv†	GGGTCACATAGACATCCCAAT	
	trYB – fw†	GGTTAATGGGGACGGACT	
	trYB – rv†	AGGAAGTTAAGATGAGGGTGG	
*trn*Y-*trn*D	trTD – fw†	TGACGATATGTCTACGCTGGT	This study
	trTD – rv*	AATCCCTGCGGGGTGTAT	
*trn*T-*trn*L	trTL – fw†	CATTACAAATGCGATGCTCT	Taberlet *et al.* (1991)
	trTL – rv†	TCTACCGATTTCGCCATATC	
ITS	ITS5 –fw†	GGAAGTAAAAGTCGTAACAAGG	White *et al.* (1990)
	ITS4 – rv†	TCCTCCGCTTATTGATATGC	

* Primer was used for amplification only.
† Primer used for both PCR amplification and sequencing.

TABLE 3. *Sequence characteristics and tree statistics of the cpDNA and ITS regions from maximum-parsimony (MP) analysis*

	cpDNA regions					Combined cpDNA regions	Combined cpDNA regions and ITS
	*psb*Z-*trn*G	*trn*Y-*trn*D	*trn*Y-*psb*M	*trn*T-*trn*L	ITS		
LAS	286–291	318–329	1028–1053	684–693	528–534	2316–2366	2844–3111
PICs*	8 (2·7 %)	12 (3·6 %)	32 (3·9 %)	19 (2·7 %)	69 (12·8 %)	71 (3·0 %)	140 (4·5 %)
TL	16	48	101	57	190	536	743
CI	0·9375	0·8958	0·6931	0·8947	0·8737	0·6250	0·6743
HI	0·0625	0·1048	0·31	0·1053	0·1263	0·3750	0·3257
RI	0·9846	0·9734	0·93	0·9757	0·9764	0·8463	0·8938
RC	0·9231	0·8720	0·6489	0·8730	0·8531	0·5252	0·6027

* Inclusive of the out-group.
LAS, length of aligned sequences; PICs, parsimony-informative characters (number and per cent); TL, tree length; CI, consistency index; HI, homoplasy index; RI, retention index; RC, rescaling consistency index.

Parsimony analysis of the ITS sequences

The aligned sequences of the ITS of the nrDNA provided the highest number of parsimony-informative characters (69; 12·8 %) of the regions used in this study, which could be attributed to an overall faster rate of base substitutions in the ITS than in the non-coding regions of the cpDNA. The ITS provided consistency and retention indices of 0·87 and 0·97, respectively (Table 3). The 50 % majority rule consensus tree from the phylogenetic analysis of DNA sequences of the ITS of 21 *Sorghum* species and *Zea mays* as an out-group species is shown in Fig. 1. Two lineages, A and E, were resolved. Lineage A was resolved with strong bootstrap support (100 %) and contained the *Eu-sorghum* species (clade B, 100 % bootstrap) and clade C with similar bootstrap support containing *S. laxiflorum* and *S. macrospermum.* The moderately supported internal clade D (61 %) contains unresolved relationships of *S. bicolor* accessions with other *Eu-sorghum* species but excludes *S. bicolor*-2 originally from Yemen. The other lineage, E, with 92 % bootstrap support contained the remaining native Australian *Sorghum* species which, except for *S. nitidum*, are contained in clade F with moderate bootstrap support (88 %; Fig. 1).

Analysis of the non-coding regions of cpDNA sequence data

The cpDNA regions, *psb*Z-*trn*G, *trn*Y-*psb*M, *trn*Y-*trn*D and *trn*T-*trn*L, revealed differences in the number of parsimony-informative characters, consistency and retention indices (Table 3). The cpDNA data show less homoplasy than the ITS data (Table 3), resulting in more fully resolved 50 % majority rule consensus trees and generally greater bootstrap values for various nodes. The *trn*Y-*psb*M spacer provided the highest number of parsimony-informative characters (32; 3·9 %). The *psb*Z-*trn*G region provided the lowest number of parsimony-informative characters (eight; 2·7 %). The *trn*T-*trn*L and *trn*Y-*trn*D intergenic spacers generated sequences that had 19 (2·7 %) and 12 (3·6 %) parsimony-informative characters, respectively. As measures of accuracy for the topologies obtained, consistency and retention indices were highest (0·94 and 0·98, respectively) for *psb*Z-*trn*G among the cpDNA regions used. The *trn*Y-*psb*M spacer had the lowest consistency index (0·69) and retention index (0·93). The 50 % majority rule consensus of 100 most-parsimonious trees is shown in Fig. 2. Lineage A is resolved and includes all the *Eu-sorghum* species, clade B with strong support (100 %), and *S. laxiforum* and *S. macrospermum* (clade C) with equal bootstrap support. The strongly supported (94 %) clade D includes all *Eu-sorghum* species but excludes *S. arundinaceum.* The strongly supported (96 %) internal clade H containing *S. almum* and *S. bicolor*-2 from Yemen excludes *S. drummondii*-2. All wild *Sorghum* species from Australian except *S. laxiflorum* and *S. macrospermum* form the second lineage (lineage J), which has very strong bootstrap

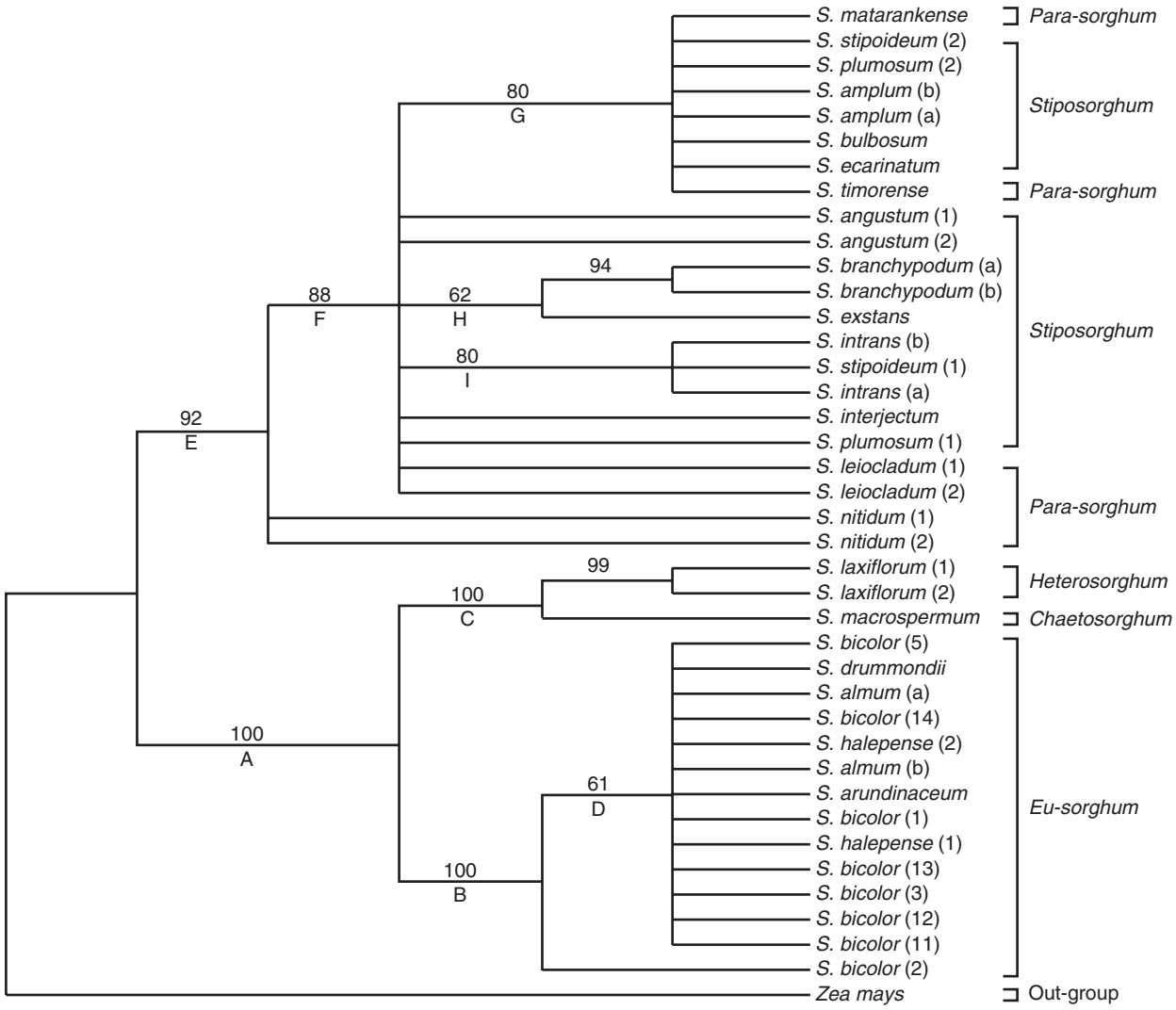

FIG. 1. Maximum-parsimony 50 % majority rule consensus tree (1000 bootstrap replicates with 100 random additions; MaxTrees = 100) generated from a phylogenetic analysis of DNA sequence data from the internal transcribed spacers of the nrDNA of 21 *Sorghum* species and *Zea mays* as an out-group species. Indels are treated as missing data. Clades are indicated by letters below the branch. Bootstrap values of >50 % are indicated above the branches.

support (100 %; Fig. 2). Clade K, with moderate bootstrap support (71 %), includes all *Stiposorghum* species and some *Parasorghum* species except *S. leiocladum* and *S. nitidum*. The internal relationships within clade K are either moderately to strongly supported by the bootstrap data (76–95 %) or remain unresolved (Fig. 2).

Combined analysis of cpDNA and ITS sequence data

The combined cpDNA and ITS sequences generated a total of 3096 characters, 140 of which (4·5 %) were parsimony-informative (Table 3). The maximum-parsimony (MP) analysis involving the combined data from the cpDNA regions and the ITS sequence data, with gaps either considered as missing values (Fig. 3) or scored as presence or absence characters (data not shown), produced two main lineages. Lineage A contains all the *Eu-sorghum* species (clade B), which includes all *S. bicolor* and their immediate wild relatives, *S.* × *almum*, *S. halepense*, *S. drummondii* and *S. arundinaceum* with 100 % bootstrap support. The other lineage, lineage J, consists of all Australian wild *Sorghum* species except *S. laxiflorum* and *S. macrospermum* with high bootstrap support (Fig. 3). *S. laxiflorum* and *S. macrospermum* not only form a single clade (C) with strong bootstrap support but are also more closely related to the *Eu-sorghum* species with 100 % bootstrap support than to other Australian wild *Sorghum* species. Within the *Eu-sorghum* section, clade D excludes *S. arundinaceum* from the rest of the species, but a subgroup comprising *S. halepense*-1, *S. drummondii*, *S. almum*, and *S. bicolor*-1, -2, -5 and -13 is formed as clade F with 99 % bootstrap support (Fig. 3). The strongly supported (94 %) clade E consists of three accessions of *S. bicolor* (-3, -11 and -14). The *S. bicolor* accessions in this clade originated from southern Africa, one from Zimbabwe (*S. bicolor*-3) and the other two from Zambia. *S. bicolor*-2, an accession from Yemen, seems to be distantly related to *S. bicolor* accessions from southern Africa but has a

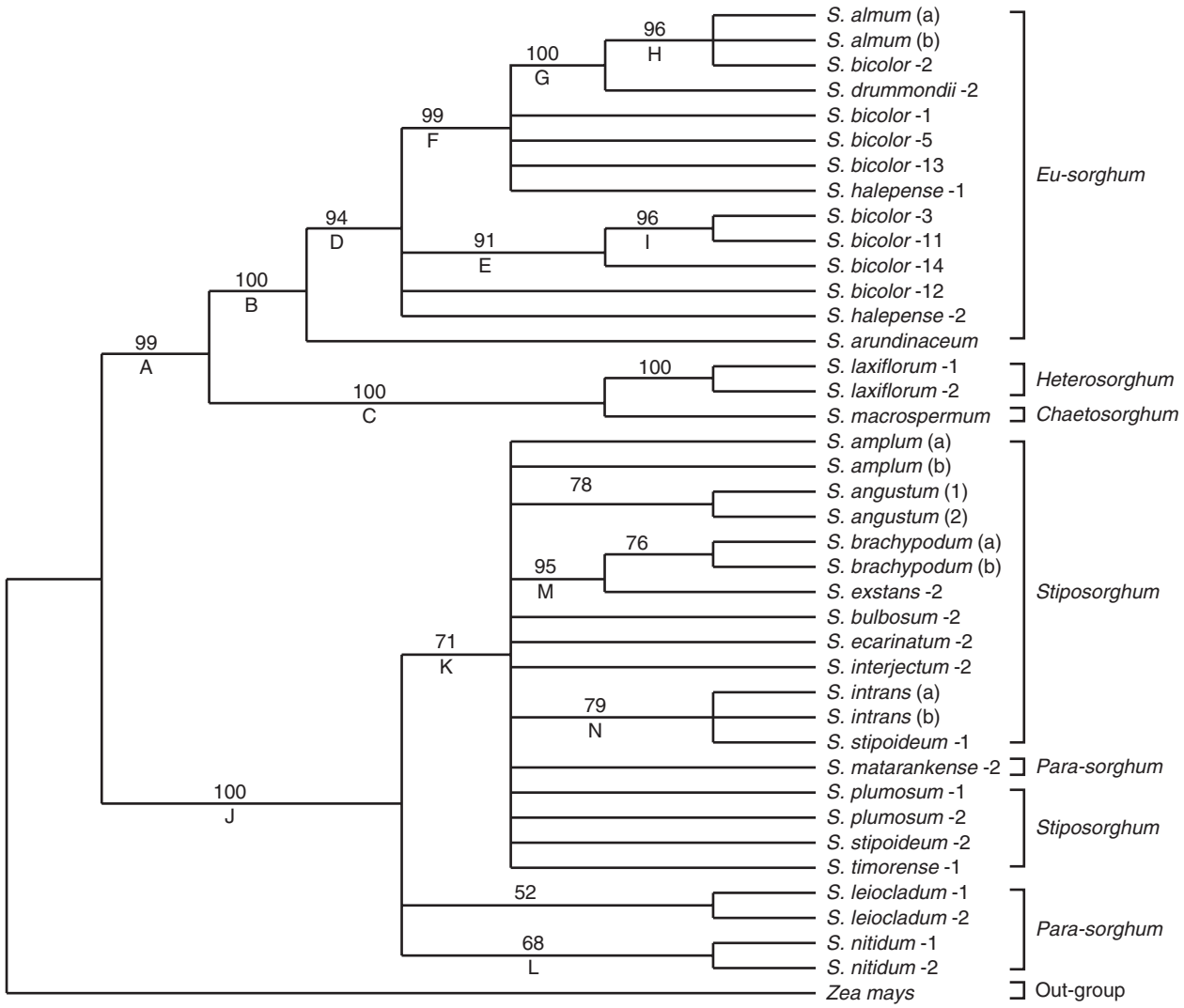

FIG. 2. Maximum-parsimony 50 % majority rule consensus tree (1000 bootstrap replicates with 100 random additions; MaxTrees = 100) from a phylogenetic analysis of DNA sequence data from the four regions of cpDNA of 21 *Sorghum* species and *Zea mays* as an out-group species. Indels are treated as missing data. Clades are indicated by letters below the branch. Bootstrap values of >50 % are indicated above the branches.

stronger association (clade H) with *S. almum* with strong bootstrap support (Fig. 3).

Stiposorghum and *Para-sorghum* form clade J with 100 % bootstrap support (Fig. 3). The internal nodes of this particular clade, however, lack strong bootstrap support. Most of the *Para-sorghum* and all of the *Stiposorghum* species form clade K with moderate bootstrap support and the two accessions of *S. nitidum* form a single clade (L) with equally moderate bootstrap support (Fig. 3). Clade M consists of *S. brachypodum* and *S. exstans* with 95 % bootstrap support. *S. intrans* and *S. stipoideum*-1 form clade N, and *S. amplum* and *S. ecarinatum* form clade O but with only moderate bootstrap support (78 %; Fig. 3).

DISCUSSION

Comparative DNA sequencing has become a widespread tool for inferring phylogenetic relationships and in systematic studies as it is relatively fast and convenient. Phylogenetic inference and elucidation of the evolutionary processes that generate biological diversity have been accomplished even at lower taxonomic levels using non-coding regions of the chloroplast genome and the ITSs of the nrDNA (Mort *et al.*, 2007; Kårehed *et al.*, 2008). In the present study, all the five cpDNA primers used successfully amplified the target regions in the *Sorghum* species. Mort *et al.* (2007) assessed the phylogenetic utility of the ITS and nine rapidly evolving cpDNA loci (including *trn*S-*trnf*M, *trn*D-*trn*T, *psb*M-*trn*D and *trn*T-*trn*L) involving six taxa sets of 13–23 taxa using published primer sequences (Shaw *et al.*, 2005). Failure of PCR amplification was reported in *Tolpis* (Asteraceae) and *Chrysosplenium* (Saxifragaceae) with the primer pair *trn*D-*trn*T. Attempts to amplify the *trn*T-*trn*L region was not successful in all the taxa used. This implies that successful amplification using published primers for some cpDNA regions of one taxon may not have universal application across taxa. In this study, *trn*Y-*psb*M provided the highest number of parsimony-informative characters followed by *trn*T-*trn*L and

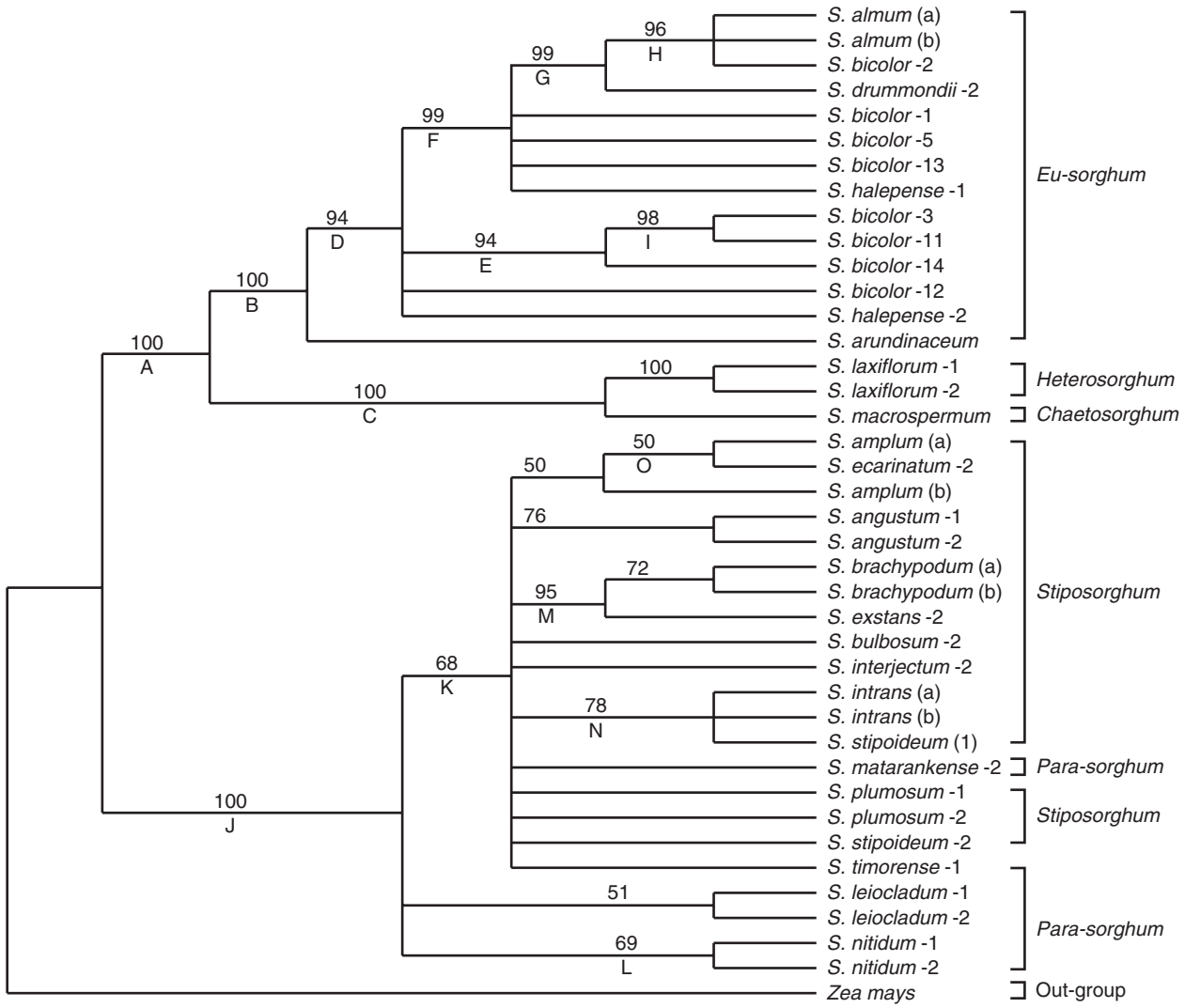

FIG. 3. Maximum-parsimony 50 % majority rule consensus tree (1000 bootstrap replicates with 100 random additions; MaxTrees = 100) from a phylogenetic analysis of DNA sequence data from the four regions of cpDNA and the internal transcribed spacers of the nrDNA of 21 *Sorghum* species and *Zea mays* as an out-group species. Indels are treated as missing data. Clades are indicated by letters below the branch. Bootstrap values of >50 % are indicated above the branches.

*trn*Y-*trn*D. Based on the potentially informative characters generated, *trnT-trnL* and *psb*M-*trn*D were identified as suitable for low taxonomic level phylogenetic studies (Shaw *et al.*, 2005). Of the cpDNA regions used in this study, *trn*Y-*psb*M, *trnT-trnL* and *trn*Y-*trn*D intergenic spacers were useful in the inference of phylogenetics at low taxonomic level in general and in the genus *Sorghum* in particular.

In the ITS analysis, all *Stiposorghum* and *Para-sorghum* species were resolved into a lineage separate from *Eu-sorghum*, *Heterosorghum* and *Chaetosorghum* species with strong bootstrap support (92 %). These results are consistent with findings based on the analysis of the ITS sequences (Sun *et al.*, 1994; Dillon *et al.*, 2001). However, in general the internal relationships between species within section are unresolved (Fig. 1). As implied and based on its utility in numerous studies, the ITS is a useful marker for resolving phylogenetic relationships at various taxonomic levels, in particular the infrageneric level. However, caution needs to be taken when analysing ITS sequence data to avoid problems resulting from concerted evolution on the rDNA arrays. Concerted evolution may homogenize different paralogous gene copies in a genome leading to the loss of all but one of the copies, i.e. different copies may be present in different organisms by chance and consequently this will create disagreement between the gene trees and species trees (Álvarez and Wendel, 2003). A fundamental requirement for historical inference based on nucleic acid or protein sequences is that the genes compared are orthologous as opposed to paralogous. However, there are inherent risks in relying exclusively on rDNA sequences for phylogenetic inferences given the 'nomadic' nature of the rDNA loci between inclusion of paralogous genes and exclusion of orthologous comparisons (Álvarez and Wendel, 2003).

The combined analysis of the cpDNA and ribosomal ITS sequence data, as when only the combined cpDNA dataset was used, resolved two major lineages (Figs 2 and 3). In

one lineage, A, the *Eu-sorghum* species form a clade B with 100 % bootstrap support. These results indicate a close association between species within the section *Eu-sorghum*. The present results are in agreement with the findings from an assessment of phylogenetic relationships among *Sorghum* taxa based on 30 allozyme loci (Morden *et al.*, 1990), which could not show clear delimitation between the *Eu-sorghum* taxa. Weedy forms of sorghum (e.g. *S. drummondii*) occur wherever cultivated sorghum and *S. arundinaceum* grow sympatrically (De Wet, 1978). Sympatric speciation, one of the theoretical models for the phenomenon of speciation, is the genetic divergence of various populations from a single parent species inhabiting the same geographical region, such that these populations become different species. However, the present study has shown emergence of two subgroups within *Eu-sorghum* with strong bootstrap support (Fig. 2). A strong phylogenetic affinity was obtained between *S. bicolor*-3, an accession from Zimbabwe, three other *S. bicolor* accessions (-11, -12 and -14) from Zambia and *S. halepense*-1, as shown in clade E. The other subgroup, clade F, contains all other *S. bicolor* accessions (-1, -2, -5 and -13; Fig. 2). Within this clade, *S. almum* is closely associated with *S. bicolor*-2, an accession from Yemen. *S. almum* is believed to be a recent fertile hybrid between *S. halepense* and *S. bicolor* (Doggett, 1970). As the chloroplast genomes are believed to display maternal inheritance in the majority of angiosperms (Mogensen, 1996; Keeling, 2004; Udall and Wendel, 2006), the present phylogenetic results suggest that *S. bicolor* could be the maternal parent of *S. almum*.

S. drummondii, commonly known as Sudan grass, is believed to be a segregate from a natural hybrid between *S. bicolor* and *S. arundinaceum* and is thought to have originated in the region from southern Egypt to the Sudan (Hacker, 1992). The cultivated species, *S. bicolor*, is allied to *S. arundinaceum*, its assumed wild progenitor (Lazarides *et al.*, 1991). This is consistent with the present results, which place *S. arundinaceum* in close relationship with *S. bicolor* with 100 % support (Fig. 3).

Various models of the origin of *S. halepense* have been suggested. Generally, the species is believed to have arisen as a segmental allotetraploid derived from the cross of two diploids ($n = 10$) species. Doggett (1970) suggested that *S. halepense* was derived from the rhizomatous perennial *S. propinquum* and the annual *S. arundinaceum*. In the allozyme variation study involving *Eu-sorghum*, *S. halepense* could not be differentiated from *S. bicolor*, suggesting that the latter was one of the parental species of *S. halepense* (Morden *et al.*, 1990). The present results (Figs 1 and 2) support the suggestion that *S. bicolor* is one of the parents of *S. halepense*.

Eu-sorghum species are closely related to *S. macrospermum* and *S. laxiflorum* with strong bootstrap support (Fig. 3), consistent with previous reports based on combined ITS1/*ndh*F/*adh*1 (Dillon *et al.*, 2007) and ITS sequence data (Sun *et al.*, 1994). This study has also revealed a very close relationship between *S. macrospermum* and *S. laxiflorum* with 100 % support (Figs 2 and 3), which suggests these species should not be classified under different sections. The close association between these two species has prompted the suggestion that *Chaetosorghum* and *Heterosorghum* be combined in a single section (Sun *et al.*, 1994; Dillon *et al.*, 2004), which is strongly supported by the present data. The ancestry of cultivated sorghum is not well resolved. Based on the ease of formation of crosses (Doggett, 1970) and chromosome morphological similarities (Gu *et al.*, 1984) within *Eu-sorghum*, it has been assumed that no other sections except *Eu-sorghum* provided the ancestral material for cultivated sorghum (van Oosterhout, 1992). However, the close association of *S. macrospermum* and *S. laxiflorum* with section *Eu-sorghum* indicates that there is strong sequence homology among them, suggesting that these species are phylogenetically closely related.

The phylogenetic relationships among the Australian wild *Sorghum* species have been described in detail (Sun *et al.*, 1994; Spangler *et al.*, 1999; Dillon *et al.*, 2001, 2004, 2007; Spangler, 2003; Price *et al.*, 2005). The internal relationships among the Australian wild sorghums are moderately well supported. *S. intrans* and *S. stipoideum* belonging to section *Stiposorghum* form a clade N with moderate support (Figs 2 and 3). These species have also been reported to be comparable in morphology and distribution (Lazarides *et al.*, 1991).

The analysis of the combined data set involving ITS and cpDNA resulted in a tree that is identical to that inferred from cpDNA alone. Similar results were obtained using the two loci on *Crassula* (Mort *et al.*, 2007). In contrast to a cpDNA-based approach, phylogenetic studies using nuclear DNA sequences have traditionally been hampered by difficulties in distinguishing between orthologous and paralogous sequences (Small *et al.*, 2004). The practice of obtaining sequence data from two or more loci that can reasonably provide independent tests of phylogeny is a proven means of avoiding well-supported but incorrect phylogenies that do not track organismal phylogeny (Mort *et al.*, 2007). Chloroplast DNA loci, which are often assumed to be uniparentally inherited and non-recombining, have been extensively used for systematics and phylogenetics. However, the rate of evolution of the cpDNA genome is slower than that of the nuclear genome. Correspondingly, the cpDNA regions that have been used for phylogenetic studies are less variable than the most extensively used nuclear loci, internal transcribed spacers of nrDNA (ITS) (Small *et al.*, 2004; Mort *et al.*, 2007). It is often difficult to obtain adequate resolution of any phylogeny of closely related taxa using few cpDNA loci due to the low number of phylogenetically informative characters (Rokas *et al.*, 2003). Hence, the practice of acquiring sequence data from several loci is a proven means of acquiring a better resolved phylogeny (Rokas and Carroll, 2005; Mort *et al.*, 2007). In the present study, the phylogeny of the genus *Sorghum* is well resolved when the combined data from ITS and four cpDNA regions were used.

Conclusions

The cpDNA regions used in this study have provided phylogenetic relationships even at low taxonomic level. The *trn*Y-*psb*M, *trn*T-*trn*L and *trn*Y-*trn*D intergenic spacers have specifically been identified to be more useful in inferring phylogenetics even at infraspecies level. The close relationship between *S. macrospermum* and *S. laxiflorum* suggest that they should not be classified under different sections and

support the proposal that sections *Chaetosorghum* and *Heterosorghum* be merged. The results also indicated that the *Eu-sorghum* species are more closely related to *S. macrospermum* and *S. laxiflorum* than to any other Australian wild *Sorghum* species. *S. almum* is more closely associated with *S. bicolor* than with *S. halepense*, its known parents. As the chloroplast genome is maternally inherited, the results suggest that *S. bicolor* is the most probable maternal parent of *S. almum*. The *S. bicolor* accessions (-3, -11 and -14) from southern Africa form a distinct and well-supported clade. *S. bicolor*-2, originally from Yemen, is distantly related to other *S. bicolor* accessions used in this study. These results may provide opportunities to use sorghum gene pools outside section *Eu-sorghum* for development and improvement of cultivated sorghum.

ACKNOWLEDGEMENTS

We thank the Nordic Genebank (now Nordgen) for financial support. We are indebted to Dr Sally Dillon of the Australian Tropical Crops and Forage Genetic Resource Centre, Biloela, Queensland, Australia, for the provision of sorghum germplasm. We would also like to thank Ms Ann-Charlotte Strömdahl at SLU, Alnarp, Sweden, for her assistance in the laboratory.

LITERATURE CITED

Álvarez I, Wendel JF. 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution* **29**: 417–434.

Baldwin BG. 1992. Phylogenetic utility of the internal transcribed spacers of the nuclear ribosomal DNA in plants. An example from the Compositae. *Molecular Phylogenetics and Evolution* **1**: 3–16.

Baldwin BG, Sanderson MJ, Porter JJ, Wojciechowski MF, Campbell CS, Donoghue MJ. 1995. The ITS region of the nuclear ribosomal DNA. A valuable source of evidence on angiosperm phylogeny. *Annals of the Missouri Botanic Gardens* **85**: 247–277.

De Wet JMJ. 1978. Systematics and evolution of *Sorghum* sect (Gramineae). *American Journal of Botany* **65**: 477–484.

De Wet JMJ, Huckay JP. 1967. The origin of *Sorghum bicolor*. II. Distribution and domestication. *Evolution* **21**: 787–802.

Dillon SL, Lawrence PK, Henry RJ. 2001. The use of ribosomal ITS to determine phylogenetic relationships within *Sorghum*. *Plant Systematics and Evolution* **230**: 97–110.

Dillon SL, Lawrence PK, Henry RJ, Ross L, Price HJ, Johnston JS. 2004. *Sorghum laxiflorum* and *S-macrospermum*, the Australian native species most closely related to the cultivated *S-bicolor* based on ITS1 and ndhF sequence analysis of 25 *Sorghum* species. *Plant Systematics and Evolution* **249**: 233–246.

Dillon SL, Lawrence PK, Henry RJ, Price HJ. 2007. *Sorghum* resolved as a distinct genus based on combined ITS1, ndhF and Adh1 analyses. *Plant Systematics and Evolution* **268**: 29–43.

Doggett J. 1970. *Sorghum*. London: Longmans, Green and Co.

Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of leaf tissue. *Phytochemical Bulletin* **19**: 11–15.

Garber ED. 1950. Cytotaxonomic studies in the genus *Sorghum*. *University of California Publications in Botany* **23**: 283–362.

Geleta M. 2007. *Genetic diversity, phylogenetics and molecular systematics of* Guizotia *Cass. (Asteraceae)*. Doctoral dissertation. Plant Protection Biology, Alnarp, Swedish University of Agricultural Sciences.

Gielly L, Taberlet P. 1994. The use of chloroplast DNA to resolve plant phylogenies – noncoding versus RBCL sequences. *Molecular Biology and Evolution* **11**: 769–777.

Gu MH, Ma TH, Liang GH. 1984. Karyotype analysis of seven species in the genus *Sorghum*. *Journal of Heredity* **75**: 196–202.

Guo Q, Huang K, Yu Y, Huang Z, Wu Z. 2006. Phylogenetic relationships of *Sorghum* and related species inferred from sequence analysis of the nrDNA ITS region. *Agricultural Sciences in China* **5**: 250–256.

Hacker JB. 1992. *Sorghum* × *drummondii* (Steud) Millsp. & Chase. In: Mannetje LT, Jones RM. eds. *Plant resources of South-East Asia. 4. Forages*. Wageningen: Pudoc Scientific Publishers, 206–208.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis programme of Windows 95/98/NT. *Nucleic Acids Symposium Series* **41**: 95–98.

Kamala V, Singh SD, Bramel PJ, Rao DM. 2002. Sources of resistance to downy mildew in wild and weedy *Sorghums*. *Crop Science* **42**: 1357–1360.

Keeling PJ. 2004. Diversity and evolutionary history of plastids and their hosts. *American Journal of Botany* **91**: 1481–1493.

Komolong B, Chakraborty S, Ryley M, Yates D. 2002. Identity and genetic diversity of the sorghum ergot pathogen in Australia. *Australian Journal of Agricultural Research* **53**: 621–628.

Kårehed J, Groeninckx I, Dessein S, Motley TJ, Bremer B. 2008. The phylogenetic utility of chloroplast and nuclear DNA markers and the phylogeny of the Rubiaceae tribe Spermacoceae. *Molecular Phylogenetics and Evolution* **49**: 843–866.

Larkin MA, Blackshields G, Brown NP, *et al.* 2007.. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948.

Lazarides M, Hacker JB, Andrew MH. 1991. Taxonomy cytology and ecology of indigenous Australian sorghums, *Sorghum* Moench, Adropogoneae, Poaceae. *Australian Systematic Botany* **4**: 591–636.

Melotto-Passarin DM, Berger IJ, Dressano K, *et al.* 2008.. Phylogenetic relationships in Solanaceae and related species based on cpDNA sequence from plastid trnE-trnT region. *Crop Breeding and Applied Biotechnology* **8**: 85–95.

Mogensen HL. 1996. The hows and whys of cytoplasmic inheritance in seed plants. *American Journal of Botany* **83**: 383–404.

Morden CW, Doebley J, Schertz KF. 1990. Allozyme variation among the spontaneous species of *Sorghum* section Sorghum (Poaceae). *Theoretical and Applied Genetics* **80**: 296–304.

Mort ME, Archibald JK, Randle CP, *et al.* 2007.. Inferring phylogeny at low taxonomic levels: utility of rapidly evolving cpDNA and nuclear ITS loci. *American Journal of Botany* **94**: 173–183.

van Oosterhout SAM. 1992. *The biosystems and ethnobotany of Sorghum bicolor in Zimbabwe*. D. phil thesis, University of Zimbabwe, Harare.

Price HJ, Dillon SL, Hodnett G, Rooney WL, Ross L, Johnston JS. 2005. Genome evolution in the genus Sorghum (Poaceae). *Annals of Botany* **95**: 219–227.

Raubeson LA, Jansen RK. 2005. Chloroplast genomes of plants. In: Henry RJ. ed. *Plant diversity and evolution: genotypic and phenotypic variation in higher plants*. Wallingford, UK: CABI Publishing, 45–68.

Rokas A, Carroll SB. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Molecular Biology and Evolution* **22**: 1337–1344.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**: 798–804.

Sharma HC, Franzmann BA. 2001. Host-plant preference and oviposition responses of the sorghum midge, *Stenodiplosis sorghicola* (Coquillett) (Dipt., Cecidomyiidae) towards wild relatives of sorghum. *Journal of Applied Ethomology* **125**: 109–114.

Shaw J, Lickey EB, Beck JT, *et al.* 2005.. The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany* **92**: 142–166.

Simmons MP, Ochoterena H. 2000. Gaps as characters in sequence-based phylogenetic analyses. *Systematic Biology* **49**: 369–381.

Small RL, Cronn RC, Wendel JF. 2004. Use of nuclear genes for phylogeny reconstruction in plants. *Australian Systematic Botany* **17**: 145–170.

Snowden JD. 1935. A classification of cultivated sorghum. *Kew Bulletin* **5**: 221–255.

Spangler RE. 2003. Taxonomy of *Sarga*, *Sorghum* and *Vacoparis* (Poaceae: Andropogoneae). *Australian Systematic Botany* **16**: 279–299.

Spangler RE, Zaitchik B, Russo E, Kellogg E. 1999. Andropogoneae evolution and generic limits in *Sorghum* (Poaceae) using *ndh*F sequences. *Systematic Botany* **24**: 267–281.

Sun Y, Skinner DZ, Liang GH, Hulbert SH. 1994. Phylogenetic analysis of *Sorghum* and related taxa using internal transcribed spacers of nuclear ribosomal DNA. *Theoretical and Applied Genetics* **89**: 26–32.

Taberlet PL, Gielly L, Pautou G, Bouvet J. 1991. Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology* **17**: 1105–1109.

Udall JA, Wendel JF. 2006. Polyploidy and crop improvement. *Crop Science* **46**: S-3–S-14.

White TJ, Bruns T, Lee S, Taylor J. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis DGM, Sninsky J, White T. eds. *PCR protocols: a guide to methods and applications*. San Diego: Academic Press, 315–322.