Original Article

Phylogenetic analysis of the genus *sorghum* based on the combined sequence data from *cpDNA* regions and the *ITS* generated strongly bootstrap 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

[2]Swedish Biodiversity Centre, Box 91, Pomonavågen 2, SE-230 53 Alnarp, Sweden

[*]Corresponding author

Email address:

Dickson Ng'uni: dickson.nguni@gmail.com

Mulatu Geleta: Mulatu.geleta.dida@ltj.slu.se

Moneim Fatih: moneim.fatih@cbm.slu.se

Tomas Bryngelsson: Tomas.Bryngelsson@ltj.slu.se

ABSTRACT

- *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 genus *Sorghum* will enhance use of the valuable agronomic traits found in the 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 analyze the phylogeny of sorghum based on maximum parsimony.
- *Key Results* Parsimony analyses of the ITS and cpDNA regions as separate or combined sequence datasets formed strongly bootstrap supported trees 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 from other *S. bicolor* accessions.
- *Conclusion 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 section. *S. almum* is closely associated with *S. bicolor* suggesting that the latter is its maternal parent considering that cpDNA is maternally inherited in angiosperms. *S. bicolor*-3, 11 and 14, from southern Africa are closely related but distantly related from *S. bicolor*-2.

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

INTRODUCTION

Sorghum Moench is highly heterogeneous which with *Cleistachne* Bentham form Sorghastrae (Garber 1950), one of the sixteen 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 twenty five 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* consists of 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 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 belong to *Chaetosorghum* and *Heterosorghum.* The section *Para-sorghum* comprises seven species. Of these, five are native to the 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 (Kamala *et al.* 2002; Komolong *et al.* 2002; Sharma and Franzmann 2001) 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 genepool.

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 noncoding chloroplast regions are phylogenetically more informative than the coding regions at lower taxonomic levels because they are under less functional constraints and evolves rapidly (Gielly and Taberlet 1994). One of the cpDNA regions, *trn*T-*trn*L used in this study was reported to possess enough phylogenetic signals for studies at lower taxonomic levels (Shaw *et al.* 2005).

The 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 that include biparental inheritance, universality of primers, intragenomic uniformity and intergenomic variability merit their utility for purposes of phylogenetic reconstruction (Baldwin *et al.* 1995). The two 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 (Dillon *et al.* 2001; Guo *et al.* 2006; Sun *et al.* 1994).

This study sought to resolve the phylogenetic relationships between the species within 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 genebank germplasm and Genbank sequence accession numbers used in this study are shown in Table 1. The germplasm accessions included the wild sorghum and some cultivated sorghum obtained from the Australian Tropical Crops Genetic resource Centre, Biloela, Australia. In addition, five accessions of *S. bicolor* and one accession of *S. arundinaceum* were obtained from the Zambian National Plant Genetic Resources Centre (ZNPGRC).

DNA extraction, PCR and sequencing

Each *Sorghum* species was represented by 1-2 accessions, except for *S. bicolor* where eleven accessions were used. Genomic DNA was extracted from fresh leaf tissues of seedlings raised in the green house approximately at two 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, Sweden).

The 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 *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 the amplification of the *trn*Y-*psb*M region. Universal primers, ITS4 and ITS5 (White *et al.* 1990), were used for the amplification and sequencing of the ITS.

The sequences of the primers and information on specific primers supplied by Eurofins MWG GmbH used in this study are given in Table 2. A GeneAMP PCR system 9700 thermocycler was used for amplification at the following temperature regime: Denaturation at 94°C for 3 min and final 7 min extension at 72°C with intervening 30 cycles of 1 min denaturing at 94°C, 1 min primer annealing temperature at 51°C and 2 min primer extension at 72°C. Successfully amplified samples were purified using the QIAquick PCR purification kit (Qiagen GmbH, Germany) and the 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) 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 consensus sequence. Since the sequences were of high quality, the forward and reverse sequences are in complete agreement, except in few cases. Such few discrepancies were resolved by repeating 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 four non-coding regions of the cpDNA were analyzed separately. In the second approach, a combined analysis included the cpDNA regions and the ITS. In these two approaches gap positions were treated as missing data. In the final approach, a combined analysis of the cpDNA regions and the ITS was undertaken, but to exploit the utility of indel positions, parsimony informative indels were coded as binary characters according to Simmons and Ochoterena (2000). *Zea mays* L. (Genbank U04796) was used as an outgroup species.

RESULTS

Sequence characteristics of the Sorghum species

The sequence characteristics and parsimony analyses based tree statistics of four non-coding regions of cpDNA and ITS have been summarized in Table 3. The aligned sequences derived from all the cpDNA regions and the ITS revealed some differences in sequence length between the *Sorghum* species. The longest sequences were obtained from the *trn*Y-*psb*M spacer ranging from 1028 (*S. drummondii*) to 1053 (*S. exstans*) to nucleotides. The eight *S. bicolor* sequences from this spacer exhibited 2-3 nucleotides differences between them. Comparatively, the *psb*Z-*trn*G spacer provided the shortest sequences that ranged between 286 (*Eu-sorghum* species) and 291 (*S. intrans*) nucleotides. The similarity in sequence length between the *Eu-sorghum* species could be attributed to the occurrence of five nucleotide

indels within the *psb*Z-*trn*G intergenic spacer. Indels of similar magnitude at corresponding positions were also observed in *S. laxiflorum* and *S. macrospermum.* Sequence length variations were also observed between *Sorghum* species in the *trn*T-*trn*L spacer, ranging in number of nucleotides from 684 (*S. arundinaceum*) to 693 (*S. leiocladum* and *S. laxiflorum*). Low sequence length differences of 2 nucleotides 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, *S. bicolor*-13 and *S. bicolor*-14 from the rest of the *S. bicolor* accessions. The sequences derived from *trn*Y-*trn*D spacer were between 318 (*S. amplum, S. angustum*) and 329 (*S. exstans*) nucleotides. The sequences obtained from the ITS showed narrow length differences between the *Sorghum* species in the range of 528-534 nucleotides. Sequence differences between Sorghum species were observed with base substitutions in the ITS1 accounting for most of the variation. The *S. bicolor* accessions exhibited sequence length differences arising from a single nucleotide indel in ITS1 region.

Parsimony analysis of the ITS sequences

The aligned sequences of the ITS of the nrDNA provided comparatively 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 chloroplast DNA. The ITS revealed the consistency and retention indices of 0.87 and 0.97 respectively (Table 3). The 50% majority rule consensus of 91 trees is shown in Figure 1. Two lineages A and E were resolved. Lineage A was resolved with strong bootstrap support (100%) that 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 that is moderately bootstrap supported (88%; Figure 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 some 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. Comparatively, 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 and percent parsimony informative characters (8; 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 on comparison of the cpDNA regions used. The *trn*Y-*psb*M spacer had the lowest consistency index (0.69) and retention index (0.93). The other non coding regions of the cpDNA had the consistency and retention indices in between these ranges. The 50% majority rule consensus of 100 trees most parsimonious trees is shown in Figure 2. Lineage A is resolved includes all the *Eu-sorghum* species, clade B with strong support (100%), *S. laxiforum* and *S. macrospermum* (clade C) with an equal bootstrap support. The strongly bootstrap supported (94%) Clade D includes all the *Eu-sorghum* species but excludes *S. arundinaceum*. The

strongly bootstrap 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) that has very strong bootstrap support (100%; Figure 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 bootstrap data (76-95%) or remain unresolved (Figure 2).

Combined analysis of cpDNA and ITS sequence data

The combined cpDNA and ITS sequences generated a total of 3096 characters of which 140 characters (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 the gaps either considered as missing values (Figure 3) or when the gaps are scored as presence or absence characters (not shown), produced two main lineages. Lineage A contains all the *Eu-sorghum* species (clade B) that includes all *S. bicolor* and their immediate wild relatives, *S.* x *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 (Figures 3). *S. laxiflorum* and *S. macrospermum* not only form the single clade (C), with strong bootstrap support but are also more closely related to the *Eu-sorghum* species with 100% bootstrap support than with 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 four accessions of *S. bicolor*-1, 2, 5 and 13 is formed as clade F with 99% bootstrap support (Figure 3). The strongly bootstrap 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 accession from Zimbabwe (*S. bicolor*-3) and other two accessions from Zambia. *S. bicolor*-2, an accession from Yemen seems to be distantly related with *S. bicolor* accessions from southern Africa but forms stronger association (clade H) with *S. almum* with strong bootstrap support (Figure 3).

Stiposorghum and *Para-sorghum* form one clade J with 100% bootstrap support (Clade J; Figure 3). The internal nodes of this particular clade, however, lack strong bootstrap support. Most of the *Para-sorghum* and all 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 (Figure 3). Clade M consists of *S. brachypodum* and *S. exstans* with 95% bootstrap support. *S. intrans* and *S. stipoideum* -1 form clade N whereas *S. amplum* and *S. ecarinatum* form clade O but with a moderate bootstrap support of 78% (Figures 3).

DISCUSSION

Comparative DNA sequencing has become a widespread tool for inferring phylogenetic relationships and 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 internal transcribed spacers of the nuclear ribosomal DNA (Kårehed *et al.* 2008; Mort *et al.* 2007). In this study, all the five cpDNA primers used in this study 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 *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 while *trn*T-*trn*L and *trn*Y-*trn*D were second and third respectively. 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 proven to be useful in the inference of phylogenetics at low taxonomic level in general and in the genus *Sorghum* in particular.

In the ITS analysis, all the *Stiposorghum* and *Para-sorghum* were resolved into a lineage separate from the *Eu-sorghum*, *Heterosorghum* and *Chaetosorghum* species with a strong bootstrap support (92%). Our results are consistent with the findings based on the analysis of the ITS sequences (Dillon *et al.* 2001; Sun *et al.* 1994). However, on the whole the internal relationships between species within section are unresolved (Figure 1). As implied and going by its utility in numerous studies, the ITS is a useful marker for resolving phylogenetic relationships at various taxonomic levels, in particular infrageneric. However, caution need to be taken when analysing ITS sequence data to avoid problems resulting from concerted evolution on the ribosomal DNA 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 (A'lvarez 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 (A'lvarez and Wendel 2003).

The combined analysis of the cpDNA and ribosomal ITS sequence data as the case when only combined cpDNA dataset was used resolved two major lineages (Figure 2 & 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*. Our 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 form(s) of sorghum and as an example, *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 geographic region, such that these populations become different species. However, our study has shown emergence of two subgroups within *Eu-sorghum* with strong bootstrap support (Figure 2). A strong phylogenetic affinity was obtained between *S. bicolor*-3, an accession from Zimbabwe and three others *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; Figure 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 (Keeling

2004; Mogensen 1996; Udall and Wendel 2006), our 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 said to have originated in the region from southern Egypt to the Sudan (Hacker 1992). The cultivated species, *S. bicolor* is allied to *S. arundinaceum*, which according to Lazarides *et al*., (1991) is the wild progenitor of *S. bicolor*. This is consistent with our results which place *S. arundinaceum* in close relationship with *S. bicolor* with 100% support (Figures 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). Our results (Figures 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 (Figures 3), consistent with the previous reports that were 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 (Figure 1), which suggests the inappropriateness of classifying these species under different sections. The close association between these two species has already

prompted a suggestion to combine *Chaetosorghum* and *Heterosorghum* into a single section (Dillon *et al.* 2004; Sun *et al.* 1994), which is strongly supported by this data. The ancestry of cultivated sorghum has not been well understood. 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 (Oosterhout van 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 elaborately described (Dillon *et al.* 2001; Dillon *et al.* 2007; Dillon *et al.* 2004; Price *et al.* 2005; Spangler 1997; Spangler 2003; Spangler *et al.* 1999; Sun *et al.* 1994). The internal relationships among the Australian wild Sorghums are moderately bootstrap supported. *S. intrans* and *S. stipoideum* (2) belonging to section *Stiposorghum* form a clade N with moderate support (Figure 1). 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 difficulties 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 proven means of avoiding obtaining well supported but incorrect phylogenies

that is not tracking 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 nuclear ribosomal DNA (ITS) (Mort *et al.* 2007; Small *et al.* 2004). 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 (Mort *et al.* 2007; Rokas and Carroll 2005). In this study, the phylogeny of the genus Sorghum is well resolved when the combined data from *ITS* and four cpDNA regions were used.

Conclusion

The cpDNA regions used in this study have shown ability to infer 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 the inappropriateness of classifying them under different sections and thus the result strongly back the proposal for merging of sections *Chaetosorghum* and *Heterosorghum*. The results also indicated that the *Eu-sorghum* species are more closely related with *S. macrospermum* and *S. laxiflorum* than with 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 strong bootstrap supported clade. *S. bicolor*-2 originally from Yemen is distantly related to other *S. bicolor* accessions in this study. These results have indications of existence of opportunities for utilization of sorghum gene pools outside the section *Eu-sorghum* for the cultivar development and improvement.

ACKNOWLEDGEMENTS

This work could not have been possible without the valuable support that has been received. We thank the Nordic Genebank (now Nordgen) for the financial support. We are indebted to Dr. Sally Dillon of Australian Tropical Crops and Forage Genetic Resource Centre LMB 1, Biloela QLD 4715 Australia for the provision of sorghum germplasm. We would also like to thank Ms. Ann-Charlotte Strömdahl at SLU, Alnarp for her assistance in the laboratory work.

LITERATURE CITED

A'lvarez 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. *Ann. Missouri Bot. Gard.* **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, Price HJ (2007) Sorghum resolved as a distinct genus based on combined ITS1, ndhF and Adh1 analyses. *Plant Systematics and Evolution* **268**, 29-43.

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.

Doggett J (1970) 'Sorghum.' (Longmans, Green and Co., London).

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. *Univ California Publ Bot* **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 x drummondii (Steud) Millsp. & Chase. In 'Plant Resources of South-East Asia. 4. Forages'. (Eds LT Mannetje, RM Jones) pp. 206-208. (Pudoc scientific Publishers: Wageningen).

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis programme of Windows 95/98/NT. *Nucl. Acids. Symp. Ser.* **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, *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, De Martin VD, Oliveira GCX, Bock R, Carrer H (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, Levsen ND, O'Leary TR, Topalov K, Wiegand CM, Crawford DJ (2007) Inferring phylogeny at low taxonomic levels: Utility of rapidly evolving cpDNA and nuclear ITS loci. *American Journal of Botany* **94**, 173-183.

Oosterhout van SAM (1992) The Biosystems and Ethnobotany of *Sorghum bicolor* in Zimbabwe. Unpublished 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 'Plant diversity and evolution: genotypic and phenotypic variation in higher plants' pp. 45-68. (CABI Publishing: Wallingford UK).

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, *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. In 'Kew Bulletin' pp. 221-255.

Spangler RE (1997) Phylogenetic relationships in the genus Sorghum (Poaceae) based on ndhF sequence data. *American Journal of Botany* **84**, 233.

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. . (Ed. DG M. Innis, J. Sninsky and T. White (editors), PCR Protocols: A Guide to Methods and Applications.) pp. 315-322. (Academic Press, San Diego, California.

Figure legends

Figure 1. The 50% majority rule consensus tree (1000 bootstrap replicates with 100 random additions; MaxTrees = 100) the parsimonious tree generated from a phylogenetic analysis DNA sequence data from the internal transcribed spacers of the nrDNA of twenty one *Sorghum* species and *Zea mays* as an outgroup species. The indels are treated as missing data. The letters below the branch denote clade. Bootstrap values greater than 50% are indicated above the branches.

Figure 2. The 50% majority rule consensus tree (1000 bootstrap replicates with 100 random additions; MaxTrees = 100) the parsimonious tree generated from a phylogenetic analysis DNA sequence data from the four cpDNA of twenty one *Sorghum* species and *Zea mays* as an outgroup species. The indels were coded as binary characters and included in the analysis. The letters below the branch denote clade. Bootstrap values greater than 50% are indicated above the branches.

Figure 3. The 50% majority rule consensus tree (1000 bootstrap replicates with 100 random additions; MaxTrees = 100) the parsimonious tree generated from a phylogenetic analysis DNA sequence data from the four cpDNA regions and the internal transcribed spacers of the nrDNA of twenty one *Sorghum* species and *Zea mays* as an outgroup species. The indels were coded as binary characters and included in the analysis. The letters below the branch denote clade. Bootstrap values greater than 50% are indicated above the branches.

Table 1: Accession identity and geographic origin of each accession of *Sorghum* species used in the study

Species	Section	Germplasm accession number	DNA sequence accession number				
			trnY-trnD	psbZ-trnG	trnY-psbM	trnT-trnL	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

The two capitalized letter superscripts at the end of the accession number denote country of origin and donor of that particular accession. A single letter means the country is a donor and origin of the accession. A=Australia; B=Burundi; E=Ethiopia; K=Kenya; T=Tanzania; Y=Yemen, Zm=Zambia and Zw=Zimbabwe.

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
psbZ-trnG	tnSM –fw[2]	TGC TTC TCC TGA TGG TTG GT	This study
	tnSM – rv[2]	GCT CGC TAC ATT GAA CTA CGC	
trnY-psbM	psBD – fw[1]	CTG TCA AGG CGG AAG CTG	This study
	psBD – rv[2]	GGG TCA CAT AGA CAT CCC AAT	
	trYB – fw[2]	GGT TAA TGG GGA CGG ACT	
	trYB – rv[2]	AGG AAG TTA AGA TGA GGG TGG	
trnY-trnD	trTD – fw[2]	TGA CGA TAT GTC TAC GCT GGT	This study
	trTD – rv[1]	AAT CCC TGC GGG GTG TAT	
trnT-trnL	trTL – fw[2]	CAT TAC AAA TGC GAT GCT CT	(Taberlet *et al.* 1991))
	trTL – rv[2]	TCT ACC GAT TTC GCC ATA TC	
ITS	ITS5 –fw[2]	GGA AGT AAA AGT CGT AAC AAG G	(White *et al.* 1990)
	ITS4 – rv[2]	TCC TCC GCT TAT TGA TAT GC	

[1] Primer was used for amplification only

[2] 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
	*psbZ-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[a]	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

[a] Inclusive of the outgroup.

LAS=Length of aligned sequences.

PICs=Parsimony informative characters (number & percent).

TL=Tree length.

CI=Consistency index.

HI=Homoplasy index.

RI=Retention index.

RC=Rescaling consistency index.

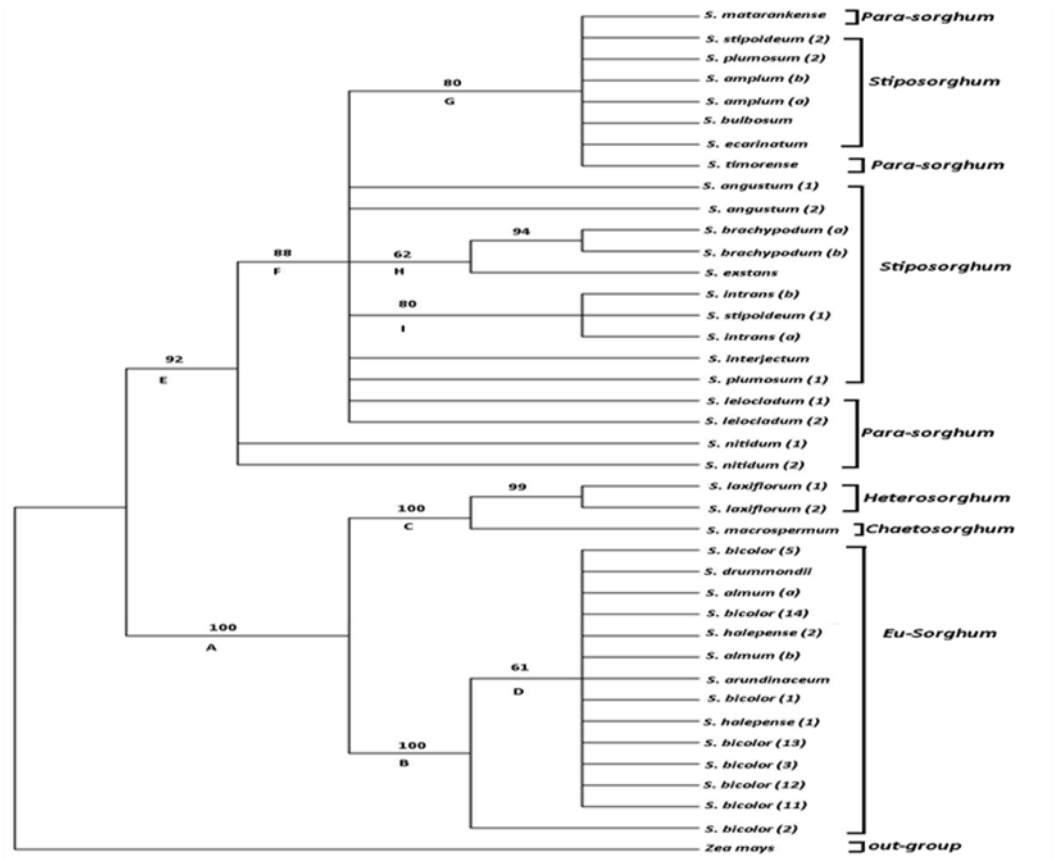

Figure 1: The 50% majority rule consensus tree of 96 most parsimonious trees (1000 bootstrap replicates with 100 random additions; MaxTrees = 100) generated from phylogenetic analysis of the ITS sequence data of twenty one *Sorghum* species and *Zea mays* as an outgroup species. The indels are treated as missing data. The letters below the branch denote clade. Bootstrap values greater than 50% are indicated above the branches.

1

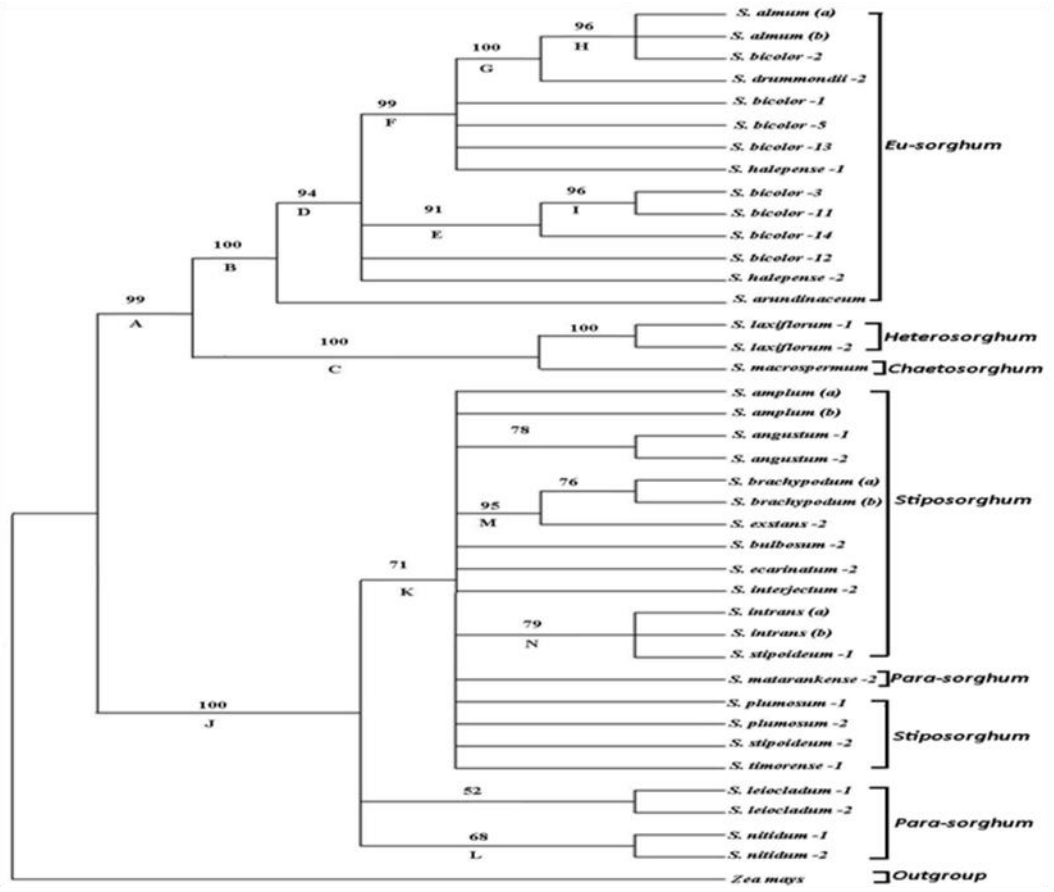

Figure 2 The 50% majority rule consensus tree of the 100 most parsimonious tree (1000 bootstrap replicates with 100 random additions; MaxTrees = 100) generated from phylogenetic analysis of sequence data from the four cpDNA regions of twenty one *Sorghum* species and *Zea mays* as an outgroup species. The indels are treated as missing data. The letters below the branch denote clade. Bootstrap values greater than 50% are indicated above the branches.

1

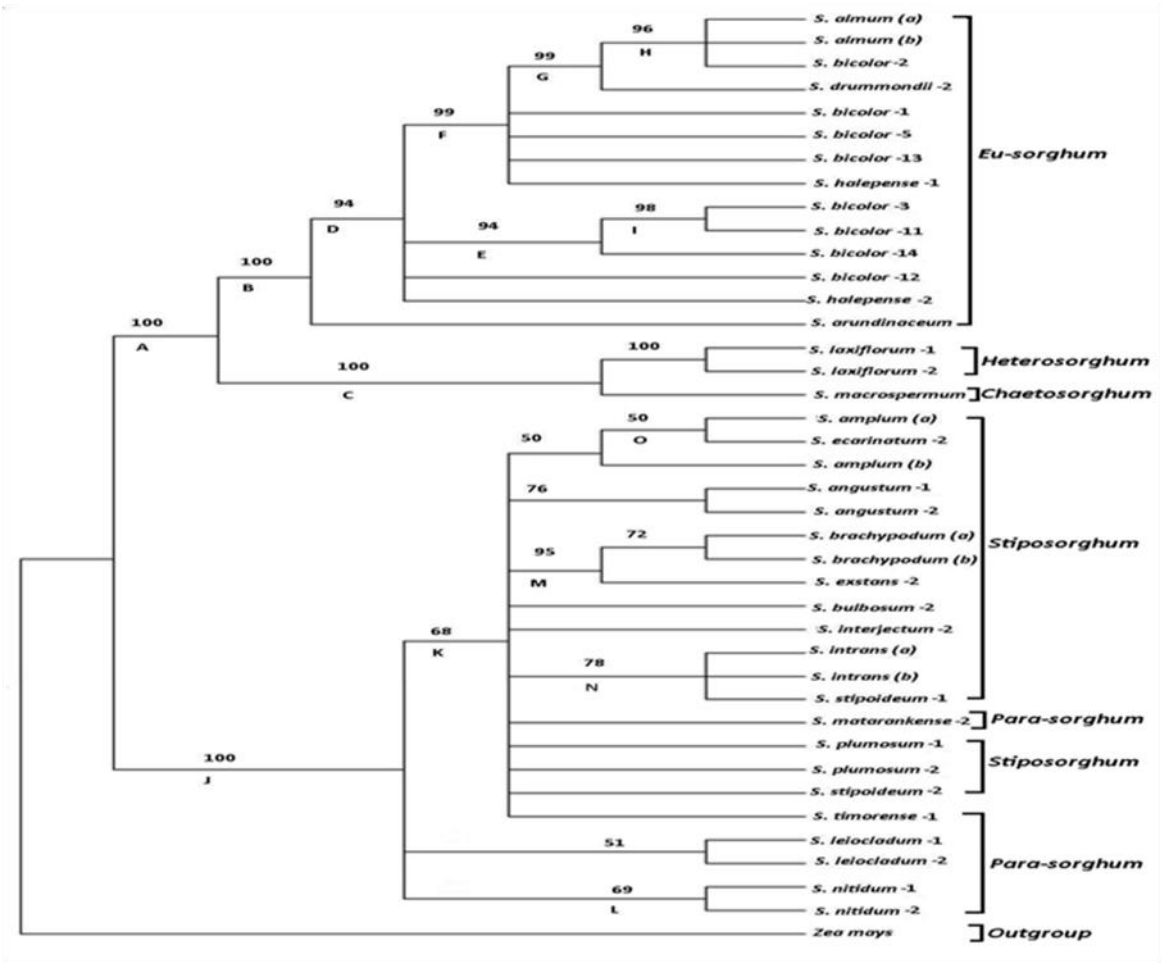

Figure 3 The 50% majority rule consensus tree of parsimonious tree (1000 bootstrap replicates with 100 random additions; MaxTrees = 100) generated from phylogenetic analysis of combined sequence data from the four cpDNA regions and the ITS of the nrDNA of twenty one *Sorghum* species and *Zea mays* as an outgroup species. The indels are treated as missing data. The letters below the branch denote clade. Bootstrap values greater than 50% are indicated above the branches.